THE AMERICAN MATHEMATICAL
MONTHLY

Volume 106, Number 8 October 1999

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper. Authors who use LaTeX are urged to use article.sty and its standard environments with no custom formatting

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues and descriptive summaries of forthcoming articles:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
DONALD ALBERS, Director of Publications

Advertising Correspondence:
Dave Riska, *driska@maa.org*

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center, *maahq@maa.org*

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Recent copies of the MONTHLY are available for purchase through the MAA Service Center, maahq@maa.org, 1-800-331-1622

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT.

Some Fundamental Control Theory I: Controllability, Observability, and Duality

William J. Terrell

1. INTRODUCTION. It is well known that a single n-th order nonhomogeneous linear differential equation is equivalent to a system of n first order linear differential equations. Specifically, an n-th order linear equation

$$y^{(n)} + k_1 y^{(n-1)} + k_2 y^{(n-2)} + \cdots + k_n y = u(t), \tag{1}$$

with real constant coefficients k_i, is equivalent, via the standard definition of the vector variable $z = [y\ y'\ y'' \ldots y^{(n-1)}]^T$, to the linear system

$$z' = Pz + du(t), \tag{2}$$

where

$$P = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & 1 \\ -k_n & -k_{n-1} & -k_{n-2} & \cdots & -k_1 \end{bmatrix} \tag{3}$$

is a companion matrix, the k_i are as in (1), and

$$d = [0\ 0 \ldots 1]^T \tag{4}$$

is the n-th standard basis vector.

What about the converse? When can a constant coefficient linear system

$$x' = Ax + bu(t), \tag{5}$$

where A is $n \times n$ and b is $n \times 1$, be transformed to (2) by a nonsingular linear transformation of the *state variable*, $z = Tx$, where T is a constant matrix? Since $z' = Tx' = (TAT^{-1})(Tx) + Tbu = (TAT^{-1})z + (Tb)u$, we are led to ask: When is there a nonsingular T such that TAT^{-1} is a companion matrix and Tb is the n-th standard basis vector?

The answer to this question is known [**8**, Chapter 2], although it seems not to be common knowledge outside the mathematical control community. A linear transformation of the state x that transforms (5) to (2) is not always possible, as can be seen by considering the diagonal system

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{6}$$

In advanced courses in dynamics the subject of normal forms achieved by coordinate transformations is an important topic. And in the literature of mathematical control theory, questions concerning alternative system representations have always been important. However, we know of no elementary differential equations text outside the control-theoretic literature that systematically addresses the question of a transformation from (5) to (2).

The primary purpose of this article is to introduce a circle of ideas in mathematical control theory. The approach is via the question of "equivalence" between n-dimensional first order linear systems like (5) and n-th order linear equations like (1). The full answer to the equivalence question introduces some of the central concepts of modern control theory. We derive some classical results concerning the important control-theoretic concepts of *controllability* and *observability*. We also consider the relationships of these concepts with other important topics in control, such as stabilization of equilibria, and linearization of nonlinear systems using coordinate change and state feedback.

- In Sections 2 and 3 we clarify the relationship between the system (5) and the equation (1) and derive a necessary and sufficient condition for equivalence.
- In Section 4 we explore the equivalence condition of Section 3 by motivating and explaining its meaning as a *controllability* condition. We then rephrase our original equivalence problem and introduce the concept of *observability*. An easy step in Section 5 then shows the algebraic duality of controllability and observability.
- In Section 6 we indicate briefly the importance of these developments to questions of asymptotic behavior such as stability.
- Finally, Section 7 briefly discusses some extensions of Sections 4–6 to the case of linear systems with multivariable input and multivariable output.

2. A SIMPLE EXAMPLE. Let us begin with a naive approach to transforming a simple example and then consider a precise definition of *linear equivalence* of systems.

Example 1. Consider the system,

$$x_1' = -2x_1 + 2x_2 + u(t) \tag{7a}$$

$$x_2' = x_1 - x_2, \tag{7b}$$

which has the form (5) with

$$A = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Differentiate (7b) and substitute from (7a) to obtain $x_2'' + 3x_2' = u(t)$. This second order equation for x_2 has the form (1) for $n = 2$, and a solution of it for $x_2(t)$ determines a function $x_1(t)$ (using $x_1 = x_2' + x_2$ from (7b)) so that system (7) is solved. Thus, system (7) can reasonably be said to be equivalent to the second order equation $y'' + 3y' = u(t)$.

Is there another second order equation of the form $y'' + k_1 y' + k_0 y = u(t)$ that is also equivalent to (7)? For example, we might try to get a second order equation for x_1. This question is handled using a precise definition of equivalence. Note that the equation $y'' + 3y' = u(t)$ has the linear system form

$$z' = \begin{bmatrix} 0 & 1 \\ 0 & -3 \end{bmatrix} z + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t). \tag{8}$$

We expect that there is a transformation from R^2 to itself that transforms our original system (7) to the form (8). Since the differential equations are linear, we expect that the transformation is linear, say $z = Tx$. Differentiation then gives: $z' = TAT^{-1}z + Tbu(t)$.

Definition 1. The system $x' = Ax + bu(t)$ is *linearly equivalent* to the system $z' = Ez + fu(t)$ if there exists a nonsingular matrix T such that

$$TAT^{-1} = E, \quad Tb = f. \tag{9}$$

Thus, (5) is equivalent to (2) if and only if there is a nonsingular T such that $TAT^{-1} = P$ and $Tb = d$, where P and d are given in (3) and (4).

In Example 1 we obtained a second order equation for the variable x_2. If we set $z_1 = x_2$, $z_2 = x_2'$, then $z_2 = x_2' = x_1 - x_2$, so a transformation demonstrating the equivalence of (7) and (8) is given by

$$T = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}.$$

The natural questions concerning existence, uniqueness, and computation of T arise. Before proceeding to answer these questions it is instructive to try to transform the following example to the form (2).

Example 2. Let A be as in Example 1, but let $b = [1\ 1]^T$. Show that this system cannot be transformed to the form (2). Hint: In Example 1 we knew P once we had the second order equation, but, in fact, we know P anyway because by similarity we know P's characteristic polynomial.

3. EQUIVALENCE AND THE COMPANION MATRIX P. System (2) is very special; we call it a *companion system* because P is a *companion matrix* defined by the characteristic polynomial $\lambda^n + k_1\lambda^{n-1} + k_2\lambda^{n-2} + \cdots + k_{n-1}\lambda + k_n$, which is the same as the characteristic polynomial of any matrix that is similar to P. Example 1 shows that system (5) may be equivalent to a companion system, and we have seen two examples of (5) that are not equivalent to a companion system, namely, Example 2 and a two-dimensional system with diagonal A having a repeated eigenvalue.

3.1 A Similarity Invariant. It is convenient to make the following definition.

Definition 2. The vector x is a *cyclic vector* for the square matrix A if the n vectors $x, Ax, \ldots, A^{n-1}x$ are linearly independent.

In (2), d is a cyclic vector for P; one way to see this is by direct calculation of

$$[d\ Pd\ P^2d\ \cdots\ P^{n-1}d] = \begin{bmatrix} 0 & \cdot & \cdot & \cdot & \cdot & 0 & 1 \\ 0 & \cdot & \cdot & \cdot & 0 & 1 & * \\ 0 & \cdot & \cdot & 0 & 1 & * & * \\ & & & \vdots & & & \\ 0 & 1 & * & * & * & * & * \\ 1 & * & * & * & * & * & * \end{bmatrix},$$

which is nonsingular. Existence of a cyclic vector for a matrix is a similarity invariant. If A is similar to P and $TAT^{-1} = P$, then A has a cyclic vector given by $T^{-1}d$.

The next proposition gives a useful condition that is equivalent to similarity between A and P.

Proposition 1. [**6**, Theorem 3.3.15] *The matrix A is similar to the companion matrix P of its characteristic polynomial if and only if the minimal and characteristic polynomials of A are identical.*

Proof: Similar matrices have the same characteristic polynomial and minimal polynomial, and the minimal polynomial of a companion matrix P is the same as its characteristic polynomial [**6**, pp. 146–147]. Thus, if A is similar to P, then the minimal polynomial and the characteristic polynomial of A must be identical.

On the other hand, if the minimal polynomial and characteristic polynomial of A are identical, then the Jordan canonical form of A must contain exactly one Jordan block for each distinct eigenvalue; the size of each Jordan block is equal to the multiplicity of the corresponding eigenvalue as a zero of the characteristic (minimal) polynomial of A. In this case, the Jordan canonical form of the companion matrix P has the same Jordan block structure as A, and hence it must be similar to A. Thus, A must be similar to P. ■

Proposition 1 makes it easy to construct examples of matrices that have (or do not have) cyclic vectors. The similarity condition holds in Examples 1 and 2, where the characteristic (and minimal) polynomial for A is $\lambda(\lambda + 3)$. We conclude that there is some other obstruction in Example 2 to an equivalence with system (2), and the obstruction must involve the b vector. Thus, the problem with transforming Example 2 is related to the way the forcing function u enters the equations. We pursue this observation in Section 4.

3.2 Uniqueness of the Transformation T. Assume that we have a nonsingular T such that $TAT^{-1} = P$ and $Tb = d$. Then $TAT^{-1}d = TAb$, and $TA^kb = TA^kT^{-1}d = (TAT^{-1})^kd = P^kd$ for all $k \geq 0$. Nonsingularity of T implies that

$$n = rank\,[d\ Pd\ P^2d\ \cdots\ P^{n-1}d] = rank\,[b\ Ab\ A^2b\ \cdots\ A^{n-1}b].$$

Moreover, T is uniquely determined by its action on the basis defined by the vectors $b, Ab, \ldots, A^{n-1}b$. Thus, we have the following uniqueness result and necessary condition.

Proposition 2. *There is at most one nonsingular linear transformation,* $z = Tx$, *taking* (5) *to the companion form* (2). *Such a* T *exists only if*

$$rank\,[b\ Ab \ldots A^{n-1}b] = n. \tag{10}$$

Example 2 is explained by this result, because in that example we have

$$[b\ Ab] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

We return to Example 2 later for additional insight. Proposition 2 also explains why we cannot get a second order equation (1) for the variable x_1 in Example 1: the second order equation for $y = z_1$ must be a *unique* linear combination of the components of x.

We now show that the criterion (10) is sufficient for there to be a nonsingular linear transformation T from (5) to (2). We also show how to construct T by a simple direct method.

3.3 The Rank Condition (10) is Sufficient for Equivalence. Referring back to Example 1, the key in transforming (5) to (2) is to identify the variable z_1 that satisfies an equivalent n-th order equation (1). Note that $z_1 = (\text{first row of } T) \cdot x$. Let τ denote the first row of T. Since $Tb = d = [0\ 0 \ldots 0\ 1]^T$ and $TA^kb = P^kd$, we must have $\tau \cdot b = 0$, $\tau \cdot Ab = 0, \ldots, \tau \cdot A^{n-2}b = 0$, and $\tau \cdot A^{n-1}b = 1$. Write this as

$$\tau[b\ Ab \ldots A^{n-1}b] = [0 \ldots 0\ 1] = d^T. \tag{11}$$

Now if we assume $rank[b\ Ab \dots A^{n-1}b] = n$, there is a unique solution for τ in (11). (Again, the crucial z_1 variable must be a *unique* linear combination of the x components.) What about the rest of T? The form of the companion system requires that

$$z_2 = (\text{second row of } T)x = z_1' = \tau \cdot x' = \tau \cdot (Ax + bu) = \tau Ax + \tau bu = \tau Ax;$$

therefore the second row of T is τA. Continuing in this way, the equations defining τ and the form of the z system imply that

$$T = \begin{bmatrix} \tau \\ \tau A \\ \tau A^2 \\ \vdots \\ \tau A^{n-1} \end{bmatrix}. \tag{12}$$

We combine this argument with Proposition 2 as follows.

Theorem 1. *The system $x' = Ax + bu(t)$ with $x \in R^n$ can be transformed to the companion system, $z' = Pz + du(t)$, by a nonsingular linear transformation, $z = Tx$, if and only if $rank[b\ Ab \dots A^{n-1}b] = n$; in this case, T is uniquely defined by* (12), *where τ is the unique solution of* (11).

Theorem 1 answers our original question. If the basic algebraic fact concerning the existence of a cyclic vector for the companion matrix of A is understood, then the situation regarding equivalence between (5) and (2) becomes transparent.

Our original question got us to this point. But there is much more involved here, if we re-examine things. Think about varying the nonhomogeneous term in (5). What if we apply different input functions $u(t)$? To what extent can this affect the solutions of the system?

We consider the question of varying the input $u(t)$ in the next section. By doing so, we obtain an analytic, control-theoretic meaning of the rank condition in Theorem 1.

4. CONTROLLABILITY. System (5) is often called a *single-input* system because the input function u is scalar-valued rather than vector-valued. We show in this section that a natural concept of *controllability* for the single-input system (5) coincides with b being a cyclic vector for A.

In an elementary differential equations course the nonhomogeneous term in (1) is considered to be a fixed, specified function of t. But we now ask: What happens with the system dynamics as we change u? More specifically, to what extent can the motion of the state vector $x(t)$ be influenced, starting from an initial state x_0 and using fairly arbitrary inputs, $u(t)$? The next definition describes a concept of complete controllability of the state. Before stating this definition, we should specify a set $\mathscr{U}$ of admissible input functions. The solutions of linear constant coefficient systems of differential equations are defined on the entire real line, and generally we want the same property for the inputs. However, for some questions, the inputs are restricted to an interval $[t_0, t_f]$. Thus, with an appropriate restriction of domain when necessary, we could consider several vector spaces of functions for the set $\mathscr{U}$, including piecewise constant, continuous, or locally integrable inputs. A real-valued function $u(t)$ is locally integrable if

$$\int_{t_1}^{t_2} |u(s)|\, ds < \infty$$

for each $t_1 < t_2$. The set of locally integrable functions is the largest vector space of inputs for which (13) makes sense; therefore we assume our inputs are locally integrable.

Definition 3. The linear system (5) is *completely controllable* if, given any x_0, $x_f \in R^n$, there exists a $t_f > 0$ and a control function $u(t)$, defined for $0 \le t \le t_f$, such that the solution to (5) with initial condition $x(0) = x_0$ satisfies $x(t_f) = x_f$.

The solution of (5) with $x(0) = x_0$ is

$$x(t) = e^{tA}\left(x_0 + \int_0^t e^{-sA} bu(s)\, ds\right), \tag{13}$$

where

$$e^{tA} = I + tA + \frac{t^2}{2!}A^2 + \cdots + \frac{t^k}{k!}A^k + \cdots.$$

By the Weierstrass M-test, this series is absolutely and uniformly convergent for $|t| \le t_f$ for any finite t_f [**10**, pp. 134–135]. The linear system (5) is completely controllable if for any given x_0, x_f there is some t_f and some locally integrable function u on $0 \le t \le t_f$ such that

$$x_f = e^{t_f A}\left(x_0 + \int_0^{t_f} e^{-sA} bu(s)\, ds\right). \tag{14}$$

It may be surprising that the solvability of (14) for arbitrary x_0, x_f is determined by a purely algebraic criterion; the explanation lies with the Cayley-Hamilton Theorem: the matrix A satisfies $p(A) = 0$, where $p(\lambda)$ is the characteristic polynomial of A. The rank condition (10) is known as the *controllability rank condition*, and the matrix $[b\ Ab \ldots\ A^{n-1}b]$ is called the *controllability matrix*, because of the next theorem.

Theorem 2. *The linear system* $x' = Ax + bu(t)$ *in* (5) *is completely controllable if and only if rank* $[b\ Ab \ldots A^{n-1}b] = n$.

Proof: By the Cayley-Hamilton theorem, for each $k \ge n$, A^k can be expressed as a linear combination of the powers $A, A^2, \ldots, A^{n-1}$. Let $\mathscr{R}$ denote the column space (range) of $[b\ Ab \ldots A^{n-1}b]$. From the definition of the matrix exponential and the fact that $\mathscr{R}$ is a closed subspace of R^n, we can conclude that the range of $e^{-sA}b$ must lie in $\mathscr{R}$ for every s. Thus, the integral on the right side of (13) must lie in $\mathscr{R}$ for all t. Take $x_0 = 0$, so the states that are reachable from the origin in finite time, by means of some input $u(t)$, must all lie within $\mathscr{R}$. Thus, if the rank condition does *not* hold, then the system is *not* completely controllable because there are states that cannot be reached from x_0. This establishes the implication: complete controllability $\Rightarrow$ *rank* $[b\ Ab \ldots A^{n-1}b] = n$.

Conversely, suppose *rank* $[b\ Ab \ldots A^{n-1}b] = n$. We must now show that (5) is completely controllable. Choose any finite time $t_f > 0$, and consider the symmetric $n \times n$ matrix

$$M = \int_0^{t_f} e^{-sA} bb^T e^{-sA^T}\, ds.$$

We first show that M is nonsingular, and then we show that nonsingularity of M implies complete controllability. So suppose that $Mv = 0$ for some v; then also

$v^T M v = 0$, and this implies that

$$0 = v^T M v = \int_0^{t_f} v^T e^{-sA} b b^T e^{-sA^T} v \, ds = \int_0^{t_f} (\psi(s))^2 \, ds,$$

where $\psi(s) = v^T e^{-sA} b$. Since $(\psi(s))^2$ is continuous and nonnegative, we conclude that $\psi(s) \equiv 0$. It follows that

$$\psi(0) = v^T b = 0, \quad \psi'(0) = -v^T A b = 0, \ldots, \psi^{(n-1)}(0) = \pm v^T A^{n-1} b = 0.$$

Therefore v is perpendicular to $\mathscr{R}$. By the rank assumption, we must have $v = 0$, and therefore M is nonsingular. Now take any two points x_0, x_f in R^n, and define the control $u(s) = b^T e^{-sA^T} x$ for $0 \le s \le t_f$, where x remains to be chosen. The solution $x(t)$ with input u and initial condition x_0 has final point x_f at time t_f provided that x can be chosen so that

$$x_f = e^{t_f A}\left(x_0 + \left(\int_0^{t_f} e^{-sA} b b^T e^{-sA^T} \, ds\right) x\right) = e^{t_f A}(x_0 + Mx).$$

But $e^{t_f A}$ is nonsingular because $(e^{t_f A})^{-1} = e^{-t_f A}$, and M is nonsingular, so $x = M^{-1}(e^{-t_f A} x_f - x_0)$. Thus, any x_0 can be steered to any x_f in time t_f, so the system is completely controllable. ■

Our proof that the controllability rank condition is sufficient for complete controllability follows an argument in [**9**, pp. 167–168]

Let us illustrate both Theorem 2 and the idea of controllability by re-examining Example 2.

Example 3. (*Example* 2 *continued*) The system is

$$x' = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u.$$

Note that $\lambda = 0$ is an eigenvalue of A, and $b = [1\ 1]^T$ is a corresponding eigenvector, so the controllability rank condition does not hold. However, A is similar to its companion matrix. Using the T computed before and $z = Tx$ we have the system

$$z' = \begin{bmatrix} 0 & 1 \\ 0 & -3 \end{bmatrix} z + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u.$$

Differentiation of the z_1 equation and substitution produces a second order equation for z_1:

$$z_1'' + 3z_1' = 3u + u',$$

which does not match (1) due to the u' term. One integration produces a first order equation

$$z_1' + 3z_1 = 3\int u \, ds + u,$$

which shows that the action of arbitrary inputs u affects the dynamics in only a one-dimensional space. The original x equations might lead us to think that u can fully affect both x_1 and x_2, but notice that the z_2 equation says that u has no affect on the dynamics of the difference $x_1 - x_2 = z_2$. Only when the initial condition for z involves $z_2(0) = 0$ can u be used to control a trajectory. That is, the inputs completely control only the states that lie in the subspace $span[b\ Ab] = span\{b\} = span[1\ 1]^T$. Solutions starting with $x_1(0) = x_2(0)$ satisfy $x_1(t) = x_2(t) = \int_0^t u \, ds + x_1(0)$. One can steer along the line $x_1 = x_2$ from any initial point to any

final point $x_1(t_f) = x_2(t_f)$ at any finite time t_f by appropriate choice of $u(t)$. On the other hand, if the initial condition lies off the line $x_1 = x_2$, then the difference $z_2 = x_1 - x_2$ decays exponentially so there is no chance of steering to an arbitrarily given final state in finite time.

When a system is completely controllable, there are generally many input functions that can implement the transfer from x_0 to x_f. This flexibility can be exploited in some applications to optimize the behavior of the system in some way, for example by minimizing a measure of the cost of carrying out the transfer. In particular, if the cost of the control action is measured by the integral

$$\int_{t_0}^{t_f} |u(s)|^2 \, ds,$$

then a control that minimizes this cost can be determined. For additional details on this problem for linear systems, see [**1**, pp. 102–105]. *Optimal control theory* is concerned with optimizing various performance indices of systems such as (5).

5. OBSERVABILITY AND DUALITY. Suppose we have a system (5) for which a certain linear combination of the state components x_i is directly measured, perhaps by some combination of instruments. We write the system and its measured output as

$$x' = Ax + bu \tag{15a}$$

$$y = c^T x, \tag{15b}$$

where c is a constant vector. The function $y(t)$ is our known output.

We now ask, when is c^T the first row of a transformation T to the companion system (2) where y is the dependent variable in (1)? If such a T exists, *then T must have the form* (12) *with* $\tau = c^T$. We must also have $Tb = d = [0 \dots 0 \ 1]^T$. Thus,

$$\operatorname{rank} T = \operatorname{rank} \begin{bmatrix} c^T \\ c^T A \\ \vdots \\ c^T A^{n-1} \end{bmatrix} = n. \tag{16}$$

In addition, since $Tb = d$, we have

$$\begin{bmatrix} c^T \\ c^T A \\ \vdots \\ c^T A^{n-1} \end{bmatrix} b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} b^T \\ b^T A^T \\ \vdots \\ b^T (A^{n-1})^T \end{bmatrix} c. \tag{17}$$

You can check that the differential equation for $z = [y \ y' \dots y^{(n-1)}]^T$ really is the companion form (2), by remembering that A satisfies its own characteristic polynomial: $A^n + k_1 A^{n-1} + \cdots + k_{n-1} A + k_n I = 0$.

Proposition 3. *There exists a nonsingular T transforming* (15a) *to companion form* (2) *with $z_1 = y = c^T x$, if and only if the rank condition* (16) *holds and* (17) *is satisfied. In this case, T is uniquely determined and is the matrix in* (16).

Note that the matrix in (16) has the same rank as the matrix $[c \, A^T c \dots (A^T)^{n-1} c]$, so that $y = c^T x$ satisfies (1) if and only if the system

$$x' = A^T x + cu \tag{18}$$

is completely controllable, by Theorem 2. Moreover, (17) shows that b^T is the first row of the transformation that takes (18) to companion form.

The connection of Proposition 3 with the system (18) leads to a fundamental duality between complete controllability and the concept of *complete observability.* The rank condition (16) is known as the *observability rank condition* and the matrix

$$\begin{bmatrix} c^T \\ c^TA \\ \vdots \\ c^TA^{n-1} \end{bmatrix}. \tag{19}$$

is called the *observability matrix* for the system (15). The rank condition implies that the system state x can be reconstructed from knowledge of y, u, and their derivatives. Here is a basic definition.

Definition 4. The system (15) is *completely observable* if, for any $x_0 = x(0)$, there is a finite time $t_f > 0$ such that knowledge of the input $u(t)$ and output $y(t)$ on $[0, t_f]$ suffices to determine x_0 uniquely.

Definition 4 could be restated using only the zero input, $u \equiv 0$. To see how a determination of x_0 is made when the observability rank condition holds, differentiate the output equation (15b) $n - 1$ times and set $t = 0$ to get

$$\begin{bmatrix} y(0) \\ y'(0) \\ \vdots \\ y^{(n-1)}(0) \end{bmatrix} = \begin{bmatrix} c^T \\ c^TA \\ \vdots \\ c^TA^{n-1} \end{bmatrix} x_0 + \text{terms dependent on } u. \tag{20}$$

Under the observability rank condition, the coefficient of x_0 is nonsingular, and we can solve for x_0 in terms of y and u and their derivatives. To illustrate, consider the system of Example 2 with output $y = [1\,0]x$. In this case, (20) gives

$$\begin{aligned} \begin{bmatrix} y(0) \\ y'(0) \end{bmatrix} &= \begin{bmatrix} c^T \\ c^TA \end{bmatrix} x_0 + \begin{bmatrix} 0 \\ c^Tb \end{bmatrix} u \\ &= \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} x_0 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u. \end{aligned}$$

Therefore the system is completely observable.

Theorem 3. *The system* (15) *is completely observable if and only if the observability rank condition* (16) *holds.*

Proof: We have already shown the sufficiency of the rank condition (16). Now assume that complete observability holds; we must show that (16) holds. For the purpose of contradiction, suppose also that the observability matrix has deficient rank; then, there is a nonzero vector v such that

$$\begin{bmatrix} c^T \\ c^TA \\ \vdots \\ c^TA^{n-1} \end{bmatrix} v = 0. \tag{21}$$

Now take $x_0 = v$ and consider the output $y = c^T e^{tA} v$ using input $u \equiv 0$. Using (21) and the definition of the matrix exponential, the series expansion for y must have all coefficients equal to zero. Thus, $y \equiv 0$; but this is also the output when $x_0 = 0$ under zero input. This contradicts the complete observability assumption. ■

Motivated by the comments following Proposition 3, we define the *dual system* of (15) to be

$$x' = A^T x + cu(t) \tag{22a}$$

$$y = b^T x. \tag{22b}$$

Then the dual of the dual of a system is the original system. With this definition we can encapsulate the discussion thus far with the following classical duality statement.

Theorem 4. *The system* (15) *is completely observable if and only if the system* (22a) *is completely controllable. The system* (15a) *is completely controllable if and only if the dual system* (22) *is completely observable.*

6. FEEDBACK, STABILIZATION, OBSERVERS, AND DUALITY. A major theme in control theory is the use of *feedback* to modify the system dynamics to achieve some desired behavior, for example to stabilize an otherwise unstable equilibrium point. In this section we indicate some advantages of an equivalence with the companion system (2) with regard to these issues. We also present one additional consequence of duality. The considerations in this section help to indicate that much can be accomplished with the control of linear systems, and thus it is desirable to have an extension of the solution of the equivalence problem involving systems (2) and (5) to the case where (5) is replaced by a single-input *nonlinear* system.

Definition 5. The linear system $x' = Ax$ is *stable* if all eigenvalues of A lie in the open left half-plane.

From the theory of linear differential equations, it is known that all solutions $x(t) \to 0$ as $t \to \infty$ if all the eigenvalues of A have negative real part. In this case, the equilibrium at the origin is asymptotically stable.

Definition 6. In system (5), *linear state feedback* is specified by $u = Kx$ where K is a real $1 \times n$ matrix. The corresponding *closed loop system* is $x' = (A + bK)x$.

Consider the companion form (2). Using linear state feedback, $u = Kx$, it is possible to assign eigenvalues arbitrarily to the resulting closed loop system, provided that the complex eigenvalues of $A + bK$ occur in complex conjugate pairs. Specifically, by setting $u = Kx = [-\alpha_n - \alpha_{n-1} \ldots - \alpha_1]x$ in (2) we get the closed loop system $z' = \tilde{P}z$, where $\tilde{P}$ has the same form as P in (3) except the last row is now $[-(k_n + \alpha_n) - (k_{n-1} + \alpha_{n-1}) \ldots - (k_1 + \alpha_1)]$. Suppose that $m_1, m_2, \ldots, m_n$ are the desired coefficients of the characteristic polynomial of the closed loop, $z' = \tilde{P}z$. With the k_i known and the m_i specified, then $\alpha_i = m_i - k_i$. Thus, the coefficients of the characteristic polynomial of $A + bK$ may be chosen so that all its roots lie in the open left half plane. And the exponential rate of

convergence of $z(t)$ to the origin can be increased, for example, by shifting the roots to the left in the complex plane.

A system that is not stable might be made stable if modified by appropriate linear feedback.

Definition 7. The linear system $x' = Ax + bu(t)$ is *stabilizable* if there exists a $1 \times n$ matrix K such that the linear system $x' = (A + bK)x$ is stable.

Theorem 5. *If $x' = Ax + bu(t)$ is completely controllable then it is stabilizable and the eigenvalues of $x' = (A + bK)x$ can be assigned arbitrarily (provided that complex eigenvalues occur in conjugate pairs) by appropriate choice of K.*

Proof: We have discussed the proof only for the special case of a companion system. By complete controllability, there is a nonsingular T with $z = Tx$ such that $z' = TAT^{-1}z + Tbu$ is a companion system. Therefore the eigenvalues of $TAT^{-1} + Tb\tilde{K}$ can be assigned as described, where $u = \tilde{K}z$ represents linear feedback for the companion system. Now the similarity

$$TAT^{-1} + Tb\tilde{K} = T(A + b\tilde{K}T)T^{-1} = T(A + bK)T^{-1}; \quad K \equiv \tilde{K}T$$

shows that the eigenvalues of $A + bK$ can be assigned by appropriate choice of feedback $u = Kx$. ■

There is a concept dual to stabilizability that involves the state-to-output interaction of system (15). We give only a very brief discussion.

Definition 8. System (15) is *detectable* if there exists an $n \times 1$ matrix L such that the system $x' = (A + Lc^T)x$ is stable.

Forming the matrix $A + Lc^T$ corresponds to output feedback given by $u = Ly = Lc^Tx$. The eigenvalues for $A + Lc^T$ are the same as those for $A^T + cL^T$, which corresponds to state feedback $u = L^T x$ in the dual system (22a). Thus *a system is detectable if and only if the dual system is stabilizable.* These are purely algebraic statements. An analytic interpretation of detectability derives from its implication that linear output feedback can be used to "detect" system trajectories asymptotically through a construction known as an *observer* system. Specifically, consider the system

$$\xi' = A\xi + Bu - L(y - c^T\xi) \tag{23}$$

where ξ is an auxiliary state that can be initialized at any vector $\xi(0)$. The auxiliary state ξ is intended to approximate the true state x, and L, a so-called "output error" gain matrix, is to be chosen so that ξ approximates x. Define the error by

$$e = x - \xi.$$

The objective is to choose L so that $e \to 0$ as $t \to \infty$. Now, subtraction of (23) from (15a) gives

$$e' = (A + Lc^T)e.$$

Theorem 6. *If the system* (15) *is detectable then L can be chosen in system* (23) *so that $e = x - \xi \to 0$ as $t \to \infty$, independently of the initial condition $\xi(0)$.*

Some comments on this construction are in order. The observer system (23) is an alternative to computing the solutions of the system (15) with a direct numerical

method. By using the known data provided by A, b, and c, together with y and u, system (23) can be simulated numerically with a guarantee that the estimated state *asymptotically* reconstructs the true system state x for (15), *independently of the initial* $\xi(0)$. If we were so lucky as to have $\xi(0) = x(0)$, then the observer equation (23) implies that $\xi(t) = x(t)$ for all t, a perfect estimate. You can think of (23) as a system with inputs u and y, and with output ξ, the desired approximation. The estimate ξ itself can be fed back to (15a), via $u = K\xi$, in place of the true state for purposes of stabilization of (15a), provided that (15a) is stabilizable. In other words, the eigenvalues of the closed loop system can be placed somewhere within the left half-plane even though only the output y is measured. Moreover, an important feature of this construction is that the controller (that is, the matrix K) and the observer (essentially the matrix L) can be designed independently while ensuring that the overall, interconnected observer/controller system is stable. To see this, use (15a) together with (23) to write the combined system for (x, ξ) as

$$\begin{bmatrix} x' \\ \xi' \end{bmatrix} = \begin{bmatrix} A & bK \\ -Lc^T & A + Lc^T + bK \end{bmatrix} \begin{bmatrix} x \\ \xi \end{bmatrix}. \tag{24}$$

We can obtain the characteristic polynomial for this system by using the following similarity transformation:

$$\begin{bmatrix} I & 0 \\ -I & I \end{bmatrix} \begin{bmatrix} A & bK \\ -Lc^T & A + Lc^T + bK \end{bmatrix} \begin{bmatrix} I & 0 \\ I & I \end{bmatrix} = \begin{bmatrix} A + bK & bK \\ 0 & A + Lc^T \end{bmatrix}.$$

Thus, the characteristic polynomial of the coefficient matrix in (24) is the product of the characteristic polynomials of $A + bK$ and $A + Lc^T$. This means that K can be designed without regard to the fact that only state estimates will be fed back, and the observer error gain L can be designed without reference to the fact that the resulting state estimates are fed back for stabilization purposes. This independent design feature is often called the *Separation Principle*.

Let us consider two examples illustrating stabilizability and detectability.

Example 4. We return to Example 2 once more, and adjoin the output equation $y = x_1$. Then the system coefficients are

$$A = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad c^T = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

This system is both stabilizable and detectable, using the feedback matrix K and observer matrix L given by

$$K = \begin{bmatrix} -2 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

because $A + bK$ then has eigenvalues -2, -3, and $A + Lc^T$ has eigenvalues -2, -1. Other choices for K and L are also possible.

Example 5. Stabilizability and detectability are not guaranteed. Consider the linear system with coefficients

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c^T = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

In this case, any 1×2 feedback matrix K produces a closed loop matrix $A + bK$ with zero as an eigenvalue; therefore, the system is not stabilizable. Also, any 2×1 matrix L yields a matrix $A + Lc^T$ with zero as an eigenvalue; therefore the system is not detectable.

7. A BRIEF NOTE ON EXTENSIONS. Let us briefly describe an extension of our discussion of the single-input systems in Sections 4–6 to the case of linear systems with multivariable input and output,

$$x' = Ax + Bu(t) \tag{25a}$$

$$y = Cx, \tag{25b}$$

where $u \in R^m$, $y \in R^p$, and thus B is $n \times m$ and C is $p \times n$. As noted before, the rank tests for controllability and observability allow for a statement of algebraic duality between these concepts, once an appropriate dual system is identified. The same principles extend to (25).

Definition 3 (complete controllability) makes sense for the m-input case; the admissible control functions are R^m-valued functions $u(t)$ such that every entry of $u(t)$ is locally integrable.

The characterization of complete controllability in Theorem 2 directly carries over to (25a) with no change in the statement. In this case, the controllability matrix $[B\ AB \ldots A^{n-1}B]$ has size $n \times nm$, and the proof proceeds as before from (13). With careful attention to the dimensions involved, the same proof carries through; the M matrix is still $n \times n$ while ψ is $1 \times m$.

Complete observability of the system (25) is defined exactly as in Definition 4, and the system (25) is completely observable if and only if the observability matrix,

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix},$$

which is now $pn \times n$, has rank n. One checks that the proof of Theorem 3 carries through as before.

The dual system of (25) is defined by

$$x' = A^T x + C^T u(t) \tag{26a}$$

$$y = B^T x, \tag{26b}$$

with matrix dimensions determined, of course, by (25). Theorem 4, which documents the algebraic duality of complete observability and complete controllability, is also valid when applied to (25) and its dual system.

The extension of Theorem 5 to the case of m-input controllable systems can be based on the single-input result: see [**13**, pp. 49 – 51] for an accessible proof that essentially reduces the m-input case to the single-input case.

Definition 7 and Definition 8 have straightforward extensions to the m-input and p-output cases. Theorem 6 on observer construction is easily seen to extend to (25); the extension is essentially notational.

Once we move to linear systems with time-dependent coefficient matrices, additional technical issues arise in any extension of observability, controllability, and their duality, although several extensions have been accomplished. For example, several useful definitions of controllability for time-varying systems are possible, all of which coalesce in the linear constant coefficient case to describe the same concept. These definitions may involve the initial time t_0, the particular initial state x_0 considered, and the time interval over which control action is to take place. The reader interested in these issues is invited to explore the references. However, let us give one further example to illustrate that time-varying systems require alternative approaches in order to describe controllability properties.

Example 6. [11] Consider the system

$$x' = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} b_1(t) \\ b_2(t) \end{bmatrix} u.$$

The general solution is

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} e^t \left(x_1(0) + \int_0^t e^{-s} b_1(s) u(s)\, ds \right) \\ e^{2t} \left(x_2(0) + \int_0^t e^{-2s} b_2(s) u(s)\, ds \right) \end{bmatrix}.$$

If b_1 and b_2 are constant, then Theorem 2 ensures that the system is completely controllable. Suppose now that $b_1(t) = e^t$ and $b_2(t) = e^{2t}$; the solution is then

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} e^t \left(x_1(0) + \int_0^t u(s)\, ds \right) \\ e^{2t} \left(x_2(0) + \int_0^t u(s)\, ds \right) \end{bmatrix}.$$

Thus, solutions that start on the line $x_2 = x_1$ at $t = 0$ always satisfy the condition $x_2(t) = e^t x_1(t)$, and from this condition we can conclude that the system is not completely controllable according to Definition 3, for if $x_1(0) = x_2(0)$, then the motion is confined to the first or third quadrant since x_1 and x_2 must have the same sign. In particular, the set of points reachable from the origin lies within those two quadrants. If we consider the controllability rank condition in a pointwise manner, that is, if we consider the following matrix for each time instant t,

$$[B(t)\ AB(t)] = \begin{bmatrix} e^t & e^t \\ e^{2t} & 2e^{2t} \end{bmatrix},$$

we obtain a nonsingular matrix. This example shows that a pointwise interpretation of the controllability rank condition of Theorem 2 does not lead to a satisfactory criterion for complete controllability of a time-varying linear system.

8. FURTHER READING. Three major themes in control theory (and in this article) involve (i) the input-to-state interaction: *controllability*, (ii) the state-to-output interaction: *observability*, and (iii) transitions between different representations of a dynamical system. We have tried to illustrate those themes in a discussion of an equivalence problem for single-input linear systems.

Two comprehensive texts that focus on time-invariant linear systems are [7] and [9]. For more on multivariable input and output, and time-varying linear systems, see [1], [2], [3], [11], [12], and [13]. The linear algebra text [4] provides a mathematician's view of some fundamental results of control-theoretic interest. The presentation of linear control theory in [13] is nicely unified around the concept of invariant subspace. Additional perspective on linear systems theory from the mathematical point of view can be obtained from [5].

ACKNOWLEDGMENT. Thanks to Dr. Robert E. Terrell at Cornell University for a helpful reading of an early version of this article.

1. S. Barnett and R. G. Cameron, *Introduction to Mathematical Control Theory*, Oxford Clarendon Press, Second Edition, 1985.
2. R. W. Brockett, *Finite Dimensional Linear Systems*, John Wiley and Sons, Inc., 1970.
3. C.-T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, 1984.
4. P. A. Fuhrmann, *A Polynomial Approach to Linear Algebra*, Springer-Verlag, 1996.
5. I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, Canadian Mathematical Society Monographs, John Wiley and Sons, 1986.
6. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
7. T. Kailath, *Linear Systems*, Prentice-Hall, Inc., 1980.
8. R. E. Kalman, P. L. Falb, and M. A. Arbib, *Topics in Mathematical Control Theory*, McGraw-Hill, 1969.
9. J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Texts in Applied Mathematics **26**, Springer, 1998.
10. W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, Inc., Second Edition, 1964.
11. W. J. Rugh, *Linear System Theory*, Prentice-Hall, Inc., Second Edition, 1996.
12. E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Second Edition, Springer-Verlag, 1998.
13. W. M. Wonham, *Linear Multivariable Control*, Third Edition, Springer-Verlag, 1985.

WILLIAM J. TERRELL received his Ph.D. in Applied Mathematics from North Carolina State University in 1990. Since 1991 he has been a member of the Department of Mathematical Sciences at Virginia Commonwealth University, Richmond, Virginia. His research interests are in the area of differential equations with special focus on systems and control theory.
Department of Mathematical Sciences, Virginia, Commonwealth University, Richmond, VA 23284-2014
wterrell@atlas.vcu.edu

The Education of a Pure Mathematician

Bruce Pourciau

Characters: Stu and Denton, mathematics majors; Integrity Jane, philosophy major and auditor from Hell; Professor Class, professor of mathematics.

Setting: a university classroom, during the first days of a course called Foundations of Analysis, taught by Professor Class.

WEDNESDAY, THE FIRST DAY

PROFESSOR CLASS Good morning. I hope everyone enjoyed a rewarding and relaxing summer. I'm pleased to see so many familiar names on my class list—except for Ms Integrity Jane.... Is she here? I'm sorry, how do you pronounce your last name?

INTEGRITY Just call me Integrity. You probably don't recognize me, because I'm a philosophy major. I'm just auditing.

PROFESSOR CLASS That's an unusual name.

INTEGRITY Tell me about it.

PROFESSOR CLASS Well, welcome Integrity and welcome everyone to the Foundations of Analysis, also known fondly around here as "The Education of a Pure Mathematician". We'll be covering logic, set theory, the real numbers, and the rest of the topics listed on the syllabus I'm handing out. As we work through these topics, we will come to appreciate the roles of definitions, axioms, and logical deduction, and learn how to read, understand, and write formal proofs. In a way, this course is a kind of ceremonial rite of passage, for in passing through it, we absorb how to think and act like pure mathematicians. Everyone has a copy of the textbook? Good. Then let's begin. Yes, Integrity?

INTEGRITY Before we get started, I'd like to ask a favor of you and the rest of the class. After doing some work in the philosophy of science last year, I signed up to audit this course because I thought the foundations of analysis would offer a paradigm for how scientists should build a field of rational and unbiased inquiry.

PROFESSOR CLASS I think you've come to the right place. If you can't find rational and unbiased inquiry in mathematics, where can you find it?

INTEGRITY Exactly. So I was wondering, what if we, all of us together, agreed on a short list of basic principles for the construction of a field of scientific inquiry. Then, as the course goes along, we can keep track of how consistent we're being with our basic principles. Would this be OK with everybody?

STU Sure, fine with me, why not.

DENTON Sounds like fun.

PROFESSOR CLASS I think it's a splendid idea. Does anyone object? No one? Looks like you have a deal, Integrity.

INTEGRITY After giving this some thought last night, I even have some possible principles to suggest.

PROFESSOR CLASS Excellent. Why don't you write them on the board, over to the side there, and we'll discuss them.

INTEGRITY All right. Here they are:

Some Possible Principles
for the Construction of a Field of Scientific Inquiry

M Know what something means before you ask if it's true.

A Build in no clearly unwarranted assumptions.

S Move from the simple to the less simple.

PROFESSOR CLASS Only three?

INTEGRITY I thought about others, as well as some variations, but these three struck me as more basic, less open to reasonable objections. For example, the variation of Principle **A**, "*Make* no clearly unwarranted assumptions", doesn't seem to work, since we often test a claim—that the earth is flat, that $x^2 = 2$ for some rational number, or whatever—by assuming its truth *temporarily* in order to study its consequences. But this is a far cry from assuming its truth *permanently*, which would build in an unwarranted assumption, turning the assumption into a *given* that could influence, even determine, the shape of further inquiry. Also, these principles obviously aren't supposed to be sufficient or anything. I'm only proposing them tentatively as rules that should be followed as we put together any rational and unbiased field of scientific inquiry. At the very least, you would think that a scientist trained in a field of inquiry that violates some of these principles ought to be aware of this fact and be able to defend the violations.

DENTON Aren't they just common sense, though?

STU Yeah, I was hoping we'd get some interesting arguments out of this, but these principles seem spineless. Who would violate them?

PROFESSOR CLASS In any case, I'll box them in and write "save" over here so they don't get erased. And speaking of obviously correct principles, let's begin our course with a few lectures devoted to formal logic.

INTEGRITY Before we do any mathematics?

PROFESSOR CLASS Sure. It seems only reasonable to review the general rules of correct thought before we apply them to the particular area of mathematics. Now

then, for us a statement will be a sentence that can be labeled true or false. In formal logic we study the truth values of complex statements that we learn how to make in precise ways from simpler statements. For example, we . . . Yes, Integrity?

INTEGRITY I'm sorry to interrupt, but I'm worried that we might already be violating Principle **A** if we continue.

PROFESSOR CLASS How's that?

INTEGRITY Well, how can we be sure that logic applies to mathematics before we do any mathematics? Wouldn't that be an unwarranted assumption? I realize it may seem odd to suggest that formal logic might not preserve truth when applied to mathematical assertions, but still . . . .

DENTON Be serious. Logic isn't up for debate. It just is.

INTEGRITY I *am* serious. Logic deals with statements, that is, sentences that must be true or false, independently of whether we can know them to be true or false. But until we understand the *meaning* of mathematical assertions, their particular character and what they're about, how can we know whether it's appropriate to assume that they are always either true or false? Putting it this way, it looks as if we're going against Principle **M** too.

PROFESSOR CLASS Formal logic goes all the way back to Aristotle. For over two thousand years, we have never found logic to conflict with our experience in the world around us. Of course this is hardly surprising, since formal logic merely sets out and studies the self-evident laws of correct reasoning. It deals with formal manipulations that preserve truth, no matter *what* the meaning of the statements. So it's prior to *every* science, including mathematics.

INTEGRITY Is it prior to quantum mechanics, for example? I remember from my philosophy of science class that some sort of "quantum logic" may fit the quantum world better than classical logic.[1] Anyway, the point is, what if the meaning of a mathematical assertion *precludes* its being regarded as always true or false independently of our knowing which? Then formal logic, and perhaps some of the procedures it sanctifies (such as the Law of the Excluded Middle) would not necessarily apply. After all, the world around us is finite, while mathematics is filled with infinite processes and structures. Isn't it unjustified at this point to assume that formal logic, which seems to work beautifully in this finite world, must necessarily also work in the infinite world of mathematics?[2]

PROFESSOR CLASS We *know* it works in mathematics. It's worked perfectly for centuries.

INTEGRITY But perhaps only because classical logic was *presupposed* in that mathematics, just as you were about to presuppose it here. How could logic *not* work in a mathematics where logic—and in particular the assumption that assertions must be true or false—was built into it from the start? How can we ask whether mathematical assertions are always true or false, until we know the *meaning* of mathematical assertions? When we use a logic that takes this "bivalence" as given, before we know what mathematical assertions are about, we are in clear violation of both Principle **A** and Principle **M**.

DENTON It seems to me that bivalence is just the formal reflection of something we all believe: that mathematical assertions somehow embody "eternal truths".

INTEGRITY I believe this too, but we should not allow this sort of "religious faith" to commit us to certain types of reasoning in mathematics, ahead of understanding the meaning of mathematical assertions.[3]

DENTON Hogwash. Nothing could be more clear than that bivalence applies to mathematical assertions. Take the Riemann Hypothesis. Either all the nontrivial complex zeros of the zeta function lie on the line $\sigma = 1/2$ or there are some that don't. The Riemann Hypothesis is either true or false, whether we can prove it or not.

INTEGRITY To repeat my mantra: you cannot know this for *sure* until you first decide on the meaning you wish to assign to mathematical assertions. That's Principle **M**. I think you are being deceived by metaphors taken literally, by talk "about complex zeros of the zeta function" that you interpret as being literally about mathematical objects that exist independently of us.[4] Your "certainty" that the Riemann Hypothesis must be true or false, independently of human knowledge, therefore rests on uncertain metaphysical speculation.

STU This feels backwards. If we throw logic out, how will we know if our thinking is correct? And how can we really throw it out anyway; it's built into our language.

INTEGRITY You're not saying that purely *linguistic* structures should determine the validity of *mathematical* structures, are you?[5] Any apparently real content in such a mathematics could turn out to be an illusion created by language. And if you accept classical logic as given, so that the idea of calling the validity of that logic into question becomes unintelligible, then you could even be *trapping* mathematics in this fantasy world: you might be fixing the legitimate modes of inquiry in ways that would prevent mathematicians from ever discovering that what had been taken as given might actually be unreliable![6] Surely this would be an intolerable situation.

Look, I know it seems awfully hypothetical—I mean, really, what are the chances that after we sort out the meaning of mathematical claims, we'll find that formal logic doesn't apply—but it's at least a possibility, isn't it?

PROFESSOR CLASS Strictly speaking, I think Integrity's point—that mathematics should precede logic—is well taken, for the transformations that preserve the truth of mathematical assertions could conceivably depend on the meaning we assign to these assertions. And strictly speaking, we do not need to formalize logic as a check on our reasoning as we go on from here. In individual cases, we can still think carefully and clearly about our assumptions and procedures to check whether our argument is correct. Common sense tells us that an argument so intricate that it cannot be checked informally, cannot be checked formally either.[7] So let's skip our description of formal logic, for the moment. We can come back to it later.

Why don't we move on then to an informal description of set theory. All of mathematics rests ultimately on set theory, in the sense that every true statement in mathematics can be reduced in principle to a statement about sets that can itself be derived from the axioms of set theory.

DENTON If sets are so basic, why not give us more than an "informal" description? This is a foundations course, after all. Give us the real stuff; we can take it.

PROFESSOR CLASS I appreciate your enthusiasm, Denton, but taking up the axioms seriously would really take a big bite out of our term. As a compromise, though, let's write down some of the axioms[8]—they're called the Zermelo-Fraenkel Axioms—and we can talk about them.

AXIOM SCHEMA OF COMPREHENSION *For any property $P(x)$ of x and any A, there is some B with $x \in B$ if and only if $x \in A$ and $P(x)$ holds.*

AXIOM OF PAIR *Given any A and B, there is a C such that $x \in C$ if and only if $x = A$ or $x = B$.*

AXIOM OF INFINITY *An inductive set exists.*

AXIOM SCHEMA OF REPLACEMENT *Suppose $P(x, y)$ is a property such that for every x there is a unique y that makes $P(x, y)$ hold. Then for every A there is some B such that for every $x \in A$ there is some $y \in B$ that makes $P(x, y)$ hold.*

INTEGRITY Shouldn't we expect axioms to be self-evident? Or at least simpler than what we derive from them?

PROFESSOR CLASS Well, these axioms become more familiar and plausible the more you work with them. This is even true when we write the axioms more rigorously. The replacement scheme axiom, for instance, could have been written this way:

Given any formula ϕ with free variables among $x, y, A, w_1, \ldots, w_n$,

$$\forall A\ \forall w_1, \ldots, w_n[\forall x \in A \exists!\ y\phi \rightarrow \exists Y\ \forall x \in A\ \exists y \in Y\phi]$$

INTEGRITY Hm. Presumably you must define the positive integers in terms of these axioms?

PROFESSOR CLASS Yes, of course.

INTEGRITY But this is an obvious violation of Principle **S**! Surely we should not define something which is already clear, natural, and immediate, such as the positive integers and mathematical induction,[9] in terms of something that is far less self-evident, such as these Zermelo-Fraenkel axioms.[10] Let's put it to a class vote. How many of you find the positive integers and mathematical induction clear, natural, and obviously correct? How many feel the same way about these axioms?

PROFESSOR CLASS Obviously I don't disagree. It's plain that we have violated Principle **S**. But most mathematicians believe there are very good reasons for starting with the Zermelo-Fraenkel axioms rather than the positive integers. These axioms have given mathematics a solid foundation for many decades. Integrity, you have another comment?

INTEGRITY Yes, I've thought of a second objection. The axioms of set theory, if taken to be true, must be regarded as meaningful, for otherwise we cross Principle **M**. But to the extent that the axioms have meaning, they appear to commit us to some sort of Platonic conception of mathematical existence. And certainly the assumption that mathematical objects enjoy this kind of metaphysical existence must be seen as an unwarranted assumption, a matter of faith rather than evidence.[11] So we have a violation of Principle **A** as well.

PROFESSOR CLASS Of course most mathematicians do find some version of Platonism congenial.[12]

INTEGRITY As do I. But should we adopt a philosophy because we find it sympathetic, because, in its congenial way, it tells us what we want to hear? Or should we look for a philosophy that provides secure support for the foundations of mathematics? How can we ever feel secure if we base mathematics on the unwarranted assumption, on our private belief, that mathematical assertions refer to some objective reality? Even if we could prove the axioms of set theory were consistent —and I've heard that we can't—we wouldn't necessarily be able to construct a model.[13]

PROFESSOR CLASS I'm beginning to agree with Integrity that taking the set theory axioms seriously leads us into conflicts with not only Principle **S** but also either Principle **A** or Principle **M**. I'm also beginning to think that these three principles are not as spineless as we thought. However, I still feel that the principles reflect common sense, and that they should guide the construction of any field of rational and unbiased scientific inquiry. So let's continue to keep track of our violations, as well as what these violations tell us about our approach to mathematical inquiry. For now why don't we content ourselves with the following informal treatment of sets . . . .

FRIDAY

PROFESSOR CLASS I'm pleased to see everyone's still with us, after the starts and stops we had on Wednesday. Today should be smoother. Normally in this course I first introduce the real numbers axiomatically and only later go through the actual construction of the reals. But I doubt that Integrity Jane would be able to suspend her disbelief long enough for me to finish the axiomatic approach; so I have decided to give the construction now.

We start by defining each individual positive integer as follows:

$$1 \equiv \{\varnothing\}, \quad 2 \equiv \{\varnothing, 1\} = 1 \cup \{1\}, \quad 3 = \{\varnothing, 1, 2\} = 2 \cup \{2\}$$

and so on. (I can see you waving, Integrity, but let me continue for a minute.) To define the *set* N of positive integers, we use the Axiom of Infinity to ensure the existence of at least one set S satisfying the following two conditions,

(a) $1 \in S$
(b) For every x, $x \in S$ implies $x \cup \{x\} \in S$,

and then let N be the intersection of all sets satisfying (a) and (b). It is then simple to see that the Peano Postulates hold for N, including of course the Principle of Induction.[14]

Now Integrity has been waving her hand and shaking her head, because I guess we can all see violations of Principle **S**.

INTEGRITY Yes. I think this development seems formal and pretty, yet somehow empty, as if the desire for empirical meaning had been lost.[15] It's a terrible violation of Principle **S**, for, again, the positive integers and mathematical induction strike us as far more immediate and clear than set theory based on the Zermelo-Fraenkel axioms. Why don't we take the positive integers and their self-evident properties as given and build up mathematics from there? Can't we do that?

PROFESSOR CLASS Perhaps we could, Integrity, but what I'm describing has been found to be a precise and elegant way to define not just the positive integers, but also the real numbers. So let's push on. At this point, the rational numbers can be defined easily and their field and order properties checked. Consult your text for the details. Now to define the real numbers, we set up an equivalence relation in the collection of all Cauchy sequences of rational numbers,

$$(a_n) \cong (b_n) \text{ if } (a_n - b_n) \text{ converges to 0 in the rationals,}$$

and then we call the resulting equivalence classes real numbers.

INTEGRITY Can I ask why you put the Cauchy sequences into equivalence classes? Why not just say a real number *is* a Cauchy sequence of rationals and that two real numbers (a_n) and (b_n) are equal provided $(a_n - b_n)$ converges to 0 in the rationals?

PROFESSOR CLASS Most mathematicians find an equality based on identity fits their Platonic sympathies better than an equality based on a convention, as you propose.[16]

STU On the other hand we don't really lose anything, apart from some unnecessary abstraction, if we drop the equivalence classes, do we? After all, no one has a problem writing $1/3 = 2/6$ to mean, not the identity of the fractions, but that an equivalence relation is satisfied.

PROFESSOR CLASS Your point is well taken, Stu and Integrity. But to continue, we can now introduce operations and an ordering and verify that our set of real numbers forms an ordered field. We'll do some of this work during our next class, on Monday. At that time we will also prove that our construction has the following basic

COMPLETENESS PROPERTY *Every bounded, nonempty set S of real numbers has a least upper bound.*

INTEGRITY And by "has" you mean ...

PROFESSOR CLASS That some real number b exists that is a least upper bound for S.

INTEGRITY I guess I'm just not clear on what meaning you are giving to "b exists".

STU It's *totally* clear! It means that there *is* such a real number b.

DENTON In other words, the set of all least upper bounds is not empty.

INTEGRITY But "has a least upper bound", "a least upper bound exists", "there is a least upper bound", "the set of all least upper bounds is not empty"—these are all synonymous expressions. They don't explain the meaning at all. Do you mean that you possess a method that specifies a b that works?

PROFESSOR CLASS I guess that would depend on what you mean by "possess", "method", and "specifies".

INTEGRITY Well, suppose we consider for simplicity a less general completeness property—that every bounded sequence of rationals has a least upper bound among the reals—and ask whether we could write a program that, given any such sequence, would compute rational approximations to the least upper bound, to within any desired tolerance.

PROFESSOR CLASS Ahh...

INTEGRITY I don't believe that we *can* write such a program. I was thinking about this last night, and it seems that applied to any infinite sequence of 0s and 1s, this program either would prove that every entry vanishes or would exhibit an entry equal to 1. Most of the well-known unresolved problems of mathematics—the Riemann Hypothesis and the Twin-Prime Conjecture among them—could be solved by such a powerful program. No program of this scope exists, and surely no one believes one will ever be written.[17]

PROFESSOR CLASS This is really very interesting, Integrity. If we could have written this program, we could have said, to give it a name, that the least upper bound exists *constructively.* But it *doesn't* exist constructively. That's your point?

INTEGRITY Yes, but what I'm really worried about is what sort of meaning you can give to "b exists" when constructive existence has been ruled out. Is there anything other than some kind of metaphysical existence left?[18]

DENTON I'm confused. What's the problem? The number b still exists; the set of all least upper bounds is still nonempty. Whether b exists *constructively* or not is only an interesting side question.

INTEGRITY The question is, what do you *mean* when you claim that "b exists". We must be clear on meaning before we can decide truth. That's Principle **M**.

PROFESSOR CLASS I suppose we mean that it is false that every x in the reals R fails to be a least upper bound for S.

INTEGRITY OK, but what then is the meaning of this *new* statement. It doesn't explain the meaning of an assertion A to say that A means that B is true, and B is true means that C is true, and so on. At some point, we have to stop and give the meaning of one of these statements on its own terms. Now whatever meaning the statement "It is false that every x in R fails to be a least upper bound for S" may have, that meaning must reside in the conditions, defined by the statement itself, that allow us to say it is true. But these conditions are plainly not conditions that we, in general, can *recognize* as being true when in fact they *are* true. We just do not have the capacity to check each x in R to see whether it fails or not. The truth conditions, where any meaning must be lurking, therefore lie beyond us, untestable, beyond our experience and our consciousness. So how could we ever be said to have acquired or formed any understanding of what it takes for such a statement to be true, that is, any understanding of the meaning of the statement?

DENTON This is getting too heavy for me.

STU Can we do some mathematics now, please?

INTEGRITY Worse, there is no way for us to manifest or communicate whatever knowledge of the meaning of this statement we might claim to possess. And surely it can't be meaningful to claim that we have knowledge of something, even implicit knowledge, if we cannot, in some circumstances at least, *reveal* that knowledge.[19] Do you see what I mean?

PROFESSOR CLASS I'm beginning to, yes. And so...

INTEGRITY And so it appears that in general the statement "b exists" has no clear meaning, unless we take existence to be constructive.

DENTON Professor?

INTEGRITY I see only two ways out, and they're both bad. On the one hand, in a brazen violation of Principle **A**, you could posit the existence of a being with infinite powers, a being who can actually *perform* the infinite, even uncountable, searches required to give (nonconstructive) assertions such as "a least upper bound b exists" some meaning, some sharable, factual content.[20] Of course, you buy this meaning at a steep metaphysical price.

DENTON Professor Class?

INTEGRITY On the other hand, as a second fall-back position, you could claim that in fact the assertions of mathematics in general *have* no meaning, that in the end doing mathematics consists of manipulating meaningless strings of symbols. But then Principle **M** forces you to give up truth as well. This strikes me as falling down, rather than falling back, for this position makes our cherished mathematics, not an inquiry into "eternal truth", but a meaningless, formal game. And I'm sure you don't see yourself as having taught generations of students a meaningless game.

DENTON Professor Class, are you all right?

INTEGRITY I hate to say it, but this whole development of analysis has a formal beauty that is hollow and meaningless at the core. It just lacks—I don't know what to call it—perhaps *integrity* is the right word.[21]

DENTON Professor!

PROFESSOR CLASS Yes, Denton, I'm fine. I was just... lost in thought, I guess, and feeling a little strange.[22] Look, I know it's not the end of the hour, but why don't we quit early today, and I'll see you on Monday.

MONDAY

PROFESSOR CLASS In light of the serious questions that have come up in our class, courtesy of Integrity's initial proposal and her persistence in carrying it out, I felt driven over the weekend to think through the attitude I have always had (you could call it the classical attitude) toward mathematical existence and mathematical truth. It seems to me that the foundations of classical analysis have fallen apart

under the gentle prodding of our three innocent-looking principles of inquiry. Either we must give up one or more of those common sense principles, or we must build up the foundations of analysis along different, perhaps more constructive, lines. Stu?

STU You know, for a while I really enjoyed sitting on the sidelines, watching the philosophical dispute here in class. But this is a mathematics class. Let's do some mathematics! A philosophical discussion concerning the nature of mathematical existence may be fun, and the position we take can certainly influence our *understanding* of mathematical assertions, but it can hardly have any relevance for the *doing* of mathematics. Proofs are still proofs; theorems are still theorems.[23]

PROFESSOR CLASS If that were the case, Stu, I would have found Integrity's questions less disturbing than I have. The truth is that your philosophical stance really *does* matter, and this has been known since Brouwer's dissertation in 1907. The classical and constructive positions on mathematical existence lead to two different kinds of mathematics: different procedures are seen as legitimate, different proofs are seen as convincing, and different assertions are seen as theorems. Certain classical statements are not even *intelligible* from the constructive point of view! Some of you have read Kuhn's *The Structure of Scientific Revolutions*? You probably thought that Kuhn's ideas couldn't apply to mathematics, but in fact I would say that the incommensurability in a shift from classical mathematics to constructive mathematics is as deep if not deeper than in any paradigm shift in physics or chemistry.[24] According to Kuhn, during a scientific revolution (and here I'm quoting) "the scientist's perception of his environment must be re-educated—in some familiar situations he must learn to see a new gestalt. After he has done so the world of his research will seem ... incommensurable with the one he had inhabited before."[25] Well that pretty much describes what happened to me this weekend, except that Kuhn's bloodless account doesn't tell you how completely disorienting and yet thrilling the process can be.

At the library on Saturday, I checked out the book *Foundations of Constructive Analysis* by Errett Bishop. Starting with the positive integers and their self-evident properties, he develops a natural and constructive version of mathematical analysis that appears to be consistent with our principles **A**, **S**, and **M**. It's what Brouwer should have done, if he'd been serious about selling his intuitionist program to the classical mathematicians.[26] Though it's out of print,[27] I received permission to copy the early chapters. Pass these copies around, please. This is your new text. Much of it will look familiar, but beware, it's a starkly different world: truth and constructive proof are one (so there's no such thing as an "unknowable truth"), mathematics precedes logic, and classical logic (in some cases) fails to preserve truth. In this world, a classically correct description of an integer—for example, that m is 0 if the Riemann Hypothesis is false and 1 if it's true—can become so much empty, meaningless talk, for to Bishop every integer can be converted in principle to decimal form by a finite, purely routine, process.[28] And every mathematical assertion ultimately reduces to a report, that if we make certain (perhaps hypothetical) computations within the positive integers, then we shall get certain results.[29] From the constructive standpoint, an assertion is true only when we are in a position to assert it, and false or absurd when being in a position to assert it would give rise to a contradiction. We can no longer say that every mathematical assertion is true or false, because clearly, for many assertions A, we are in no position to assert A nor in any position to assert that A can never be asserted. And we can no longer

rely on proofs by contradiction, because knowing that "It is absurd that A is absurd" does not imply that we can assert A, for it does not imply that can necessarily effect the construction required to assert A.[30]

Well, everybody ready? We're starting the course over. We begin with the positive integers: $1, 2, 3, \ldots$

NOTES

1 See [**1**, p. 320].

2 "Concerning the grounds for accepting logical laws ... any 'justification' of such laws can be given only in terms of the adequacy of the language in which they are [embedded] to the specific tasks for which that language is employed .... Under the pressure of factual observation and norms of convenience familiar language habits may come to be revised; [so] the acceptance of logical principles as canonical need be neither on arbitrary grounds nor on grounds of their allegedly inherent authority, but on the ground that they effectively achieve certain postulated ends." Ernest Nagel in [**1**, p. 320]

3 "It seems to me that to clarify the sense of your [claim] you must again refer to metaphysical concepts: to some world of mathematical things existing independently of our knowledge .... But I repeat that mathematics ought not to depend on such notions as these. In fact all mathematicians ... are convinced that in some sense mathematics bears upon eternal truths, but when trying to define precisely this sense, one gets entangled in a maze of metaphysical difficulties. The only way to avoid them is to banish them from mathematics." A. Heyting [**11**, p. 3]

4 The contemporary mathematician's use of language "seems to force us to choose between what are in fact two metaphorical descriptions of the manner in which pure mathematical knowledge is acquired: discovery or creation. And it strongly compels us to accept that the 'correct' answer is 'discovery' and not 'creation.' ... one will then be drawn almost immediately into a completely Platonistic conception .... However ... so long as we do not fall for the idea that talk 'about statements' and 'about answers' must be taken literally as being about 'things' that stand in a certain relationship to us ... there is no choice to be made." Gabriel Stolzenberg [**17**, p. 244]
"We can, after all, ask: What does it mean for a set to exist if it can perhaps never be defined? It seems clear that this existence can only be a manner of speaking, which can only lead to purely formal propositions—perhaps made up of very beautiful *words*—about objects *called* sets. But most mathematicians want mathematics to deal, ultimately, with performable computing operations and not to consist of formal propositions about objects called this or that." Thoralf Skolem in [**10**, p. 300]

5 "Suppose that a ... mathematical construction has been carefully described by means of words, and then, the introspective character of the mathematical construction being ignored for a moment, its linguistic description is considered by itself and submitted to a linguistic application of classical logic. Is it then always possible to perform a languageless mathematical construction finding its expression in the logico-linguistic figure in question? A careful examination reveals that ... with regard to the principle of the excluded third, except in special cases, the answer is in the negative." L. E. J. Brouwer in [**9**, p. 236–7]

6 [**17**, p. 225]

7 [**4**, p. 5]

8 Taken from [**12**]

9 "[Mathematical induction], inaccessible to analytic proof and to experiment, is the exact type of the *a priori* synthetic intuition. ... Why then is this view imposed upon us with such irresistible weight of evidence? It is because it is only the affirmation of the power of the mind which knows it can conceive of the indefinite repetition of the same act, when the act is once possible. The mind has a direct intuition of this power ... ." H. Poincaré in [**1**, p. 388]

10 "Set-theoreticians are usually of the opinion that the notion of integer should be defined and that the principle of mathematical induction should be proved. But it is clear that we cannot define and prove ad infinitum; sooner or later we come to something that is not further definable or provable. Our only concern, then, should be that the initial foundations be something immediately clear, natural, and not open to question. This condition is satisfied by the notion of integer and by inductive inferences, but it is decidedly not satisfied by set-theoretic axioms of the type of Zermelo's or anything else of that kind; if we were to accept the reduction of the former notions to the latter, the set-theoretic notions would have to be simpler than mathematical induction, and reasoning with them less open to question, but this runs entirely counter to the actual state of affairs." Skolem in [**10**, p. 299]

11 The axioms of set theory "if interpreted as meaningful statements, necessarily presuppose a kind of Platonism, which cannot satisfy any critical mind and which does not even produce the conviction that they are consistent." K. Gödel, as quoted in [**9**, p. 99], in 1933.

12 "It seems to me that no philosophy can possibly be sympathetic to a mathematician which does not admit, in one manner or another, the immutable and unconditional validity of mathematical truth. Mathematical theorems are true or false; their truth or falsity is absolute and independent of our knowledge of them. In some sense, mathematical truth is part of objective reality." G. H. Hardy in [**8**, p. 1246].

13 "Suppose we have in some way proved, without thinking of any mathematical interpretation, that a logical system constructed from some linguistic axioms is non-contradictory.... If we then also find a mathematical interpretation of these axioms, does it follow... that such a mathematical system *exists*? But that has never been proved... ." Brouwer in [**19**, p. 266–7]

14 Taken from [**2**, p. 79]

15 "[A] feeling for reality... ought to be preserved in even the most abstract studies." Bertrand Russell [**16**, p. 169]

16 [**3**, p. 13], [**4**, p. 12], [**5**, p. 15]

17 [**3**, p. 4–5], [**5**, p. 7–8]

18 "If 'to exist' does not mean 'to be constructed,' it must have some metaphysical meaning. It cannot be the task of mathematics to investigate this meaning or to decide whether it is tenable or not. We have no objection against a mathematician privately admitting any metaphysical theory he likes, but Brouwer's [and more generally the constructive] program entails that we study mathematics as something simpler, more immediate than metaphysics, [as something where] 'to exist' must be synonymous with 'to be constructed.'" Heyting [**11**, p. 2]

19 [**7**, p. 225]

20 "Classical mathematics concerns itself with operations that can be carried out by God.... You may think that I am making a joke... by bringing God into the discussion. This is not true. I am doing my best to develop a secure philosophical foundation, based on meaning rather than formalistics, for current mathematical practice. The most solid foundation available at present seems to me to involve the consideration of a being with non-finite powers—call him God or whatever you will—in addition to the powers possessed by finite beings." Errett Bishop [**4**, p. 9]

21 "When I attempt to express in positive terms that quality in which contemporary mathematics is deficient,... I keep coming back to the term 'integrity.' Not the integrity of an isolated formalism that prides itself on the maintenance of its own standards of excellence, but an integrity that seeks common ground in the researches of pure mathematics, applied mathematics, and ... physics; that seeks to extract the maximum meaning from each new development; that is guided primarily by considerations of content rather than elegance and formal attractiveness; that sees to it that the mathematical representation of reality does not degenerate into a game... ." Bishop [**4**, p. 4]

22 "To anyone who starts off inside the contemporary mathematician's belief system, the discovery that an entire component of the 'reality' of one's experience is produced by acts of acceptance as such in the domain of language use is not merely illuminating. In a literal sense, it is shattering: Once a mathematician has seen that his perception of the 'self-evident correctness' of the law of excluded middle is nothing more than the linguistic equivalent of an optical illusion, neither his practice of mathematics nor his understanding of it can ever be the same." Stolzenberg [**17**, p. 268]

23 "All philosophical differences... ought not to affect the detail of mathematics, but only the interpretation. Mathematics would be in a bad way if it could not proceed until [the philosophical disputes] had been settled." Russell in 1906, the year before Brouwer's dissertation provided evidence that a constructive position on mathematical existence changes the face of mathematics, as quoted in [**14**, p. 131–132]

24 Read [**15**] and [**17**]

25 [**13**, p. 112]

26 See [**15**]

27 Bishop's book has been born again in a somewhat expanded and altered form in [**5**]

28 [**4**, p. 8]

29 [**3**, p. 3] or [**5**, p. 5]

30 For much more detail on the consequences of the constructive standpoint, read [**3**], [**4**], [**5**], [**6**], or [**11**].

ACKNOWLEDGMENTS. This article was written in the Spring of 1998, at the end of a very enjoyable academic year spent sheltered and supported by the Dibner Institute for the History of Science and Technology at MIT. I thank Mark Steiner and Gabriel Stolzenberg for their thoughtful comments.

1. Benacerraf, Paul and Hilary Putnam (eds), *Philosophy of Mathematics: Selected Readings*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.
2. Binmore, K. G., *The Foundations of Analysis: A Straightforward Introduction, Book 1: Logic, Sets and Numbers*, Cambridge University Press, Cambridge, 1980.
3. Bishop, Errett, *Foundations of Constructive Analysis*, McGraw-Hill Book Company, New York,1967.
4. Bishop, Errett, Schizophrenia in Contemporary Mathematics, in *Errett Bishop: Reflections on Him and His Research*, Murray Rosenblatt (ed.), American Mathematical Society, Providence, 1985, pp. 1–32.
5. Bishop, Errett and Douglas Bridges, *Constructive Analysis*, Springer-Verlag, New York, 1985; an outgrowth of *Foundations of Constructive Analysis* by Errett Bishop, 1967.
6. Bridges, Douglas and Fred Richman, *Varieties of Constructive Mathematics*, Cambridge University Press, Cambridge, England, 1987.
7. Dummett, Michael, *Truth and Other Enigmas*, Harvard University Press, Cambridge, 1978.
8. Ewald, William (ed.), *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, Vol. II, Clarendon Press, Oxford, 1996.
9. George, Alexander (ed.), *Mathematics and the Mind*, Oxford University Press, Oxford, 1994.
10. Heijenoort, Jean van (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic 1879–1931*, Harvard University Press, Cambridge, 1967.
11. Heyting, A. *Intuitionism: An Introduction*, North-Holland Publishing Company, Amsterdam, 1971.
12. Hrbacek, Karel and Thomas Jech, *Introduction to Set theory*, Second Edition, Marcel Dekker, Inc., New York, 1984.
13. Kuhn, Thomas S., *The Structure of Scientific Revolutions*, Second Edition, Enlarged, The University of Chicago Press, Chicago, 1970.
14. Moore, Gregory H., *Zermelo's Axiom of Choice: Its Origins, Development, and Influence*, Springer-Verlag, New York, 1982.
15. Pourciau, Bruce, Intuitionism as a Kuhnian Revolution in Mathematics, preprint.
16. Russell, Bertrand, *Introduction to Mathematical Philosophy*, George Allen and Unwin, Ltd., London, 1919.
17. Stolzenberg, Gabriel, Can an Inquiry into the Foundations of Mathematics Tell Us Anything Interesting about Mind?, in *Psychology and Biology of Language and Thought: Essays in Honor of Eric Lenneberg*, Academic Press, New York, 1978, pp. 221–269.
18. Stolzenberg, Gabriel, Review of Errett Bishop's *Foundations of Constructive Analysis*, in *Bull. Amer. Math. Soc.* 76 (1970) 301–323.
19. van Stigt, Walter P., *Brouwer's Intuitionism*, Studies in the History and Philosophy of Science, Vol. 2, North-Holland, Amsterdam, 1990.

BRUCE POURCIAU holds a B.A. from Brown and his Ph.D. from UC San Diego (under Hubert Halkin). In 1976, as an analyst specializing in optimization theory, he joined the Department of Mathematics at Lawrence University, overlooking the serene Fox River in Appleton, Wisconsin. This article emerged at the end of a year spent indulging his more recent attractions, to the mathematical foundations of Newton's *Principia* and the philosophy of mathematics, especially intuitionism, at the Dibner Institute for the History of Science (MIT), overlooking the serene Charles River. Outside of mathematics, when he isn't playing tennis, photographing nature, listening to Mozart, or reading mysteries, you will find him involved with his most important interests—his wife, Nancy, and their three children.

Lawrence University, Appleton, WI 54912
bruce.h.pourciau@lawrence.edu

Multivariable Calculus and the Plus Topology

Daniel J. Velleman

Among the most subtle concepts in multivariable calculus are the concepts of continuity and differentiability of functions of two (or more) variables. These concepts are designed to tell us about the local behavior of a function near a point. Since "local" is defined by reference to the standard topology on $\mathbf{R}^2$, the definitions of continuity and differentiability must take into account the fact that a neighborhood of a point in this topology includes nearby points in all directions, not just the coordinate directions. As a result, these definitions involve limits in which a point (x, y) approaches a point (a, b), and such limits cannot be understood in terms of limits in which the variables x and y approach the limits a and b separately. This explains why, for example, differentiability of a function of two variables is not the same as existence of the two first partial derivatives.

But now suppose we are interested in studying the partial derivatives of a function. Since the partial derivatives are defined in terms of limits with respect to the independent variables separately, they cannot be thought of as giving us information about the local behavior of the function near a point—at least, not if "local" is defined by reference to the standard topology. But what if we use a different topology? Is there some topology on $\mathbf{R}^2$ that is appropriate for the study of partial derivatives, in the same way that the standard topology is appropriate for the study of continuity and differentiability? My purpose in this paper is to show that there is such a topology, and that the study of this topology can shed light on some of the subtleties of multivariable calculus.

The standard topology on $\mathbf{R}^2$ is defined by reference to ε-balls, where for any $\varepsilon > 0$ and any point $(a, b) \in \mathbf{R}^2$, the *ε-ball centered at* (a, b) is defined to be the set

$$B_\varepsilon(a, b) = \left\{(x, y) \in \mathbf{R}^2 \middle| \sqrt{(x-a)^2 + (y-b)^2} < \varepsilon\right\}.$$

We define the *ε-plus centered at* (a, b) to be the set

$$+_\varepsilon(a, b) = \{(x, b) \in \mathbf{R}^2 \mid |x - a| < \varepsilon\} \cup \{(a, y) \in \mathbf{R}^2 \mid |y - b| < \varepsilon\}.$$

Of course, the reason for the name is that the set $+_\varepsilon(a, b)$ looks like a plus sign centered at (a, b), with "radius" ε; see Figure 1. We say that a set $U \subseteq \mathbf{R}^2$ is *plus-open* if for every $(a, b) \in U$ there is some $\varepsilon > 0$ such that $+_\varepsilon(a, b) \subseteq U$. It is easy to verify that the plus-open sets form a topology on $\mathbf{R}^2$, which we will call the *plus topology*. Clearly every open set is plus-open, but there are plus-open sets that are not open. For example, the set

$$A = \{(0, 0)\} \cup \{(x, y) \in \mathbf{R}^2 \mid |y| > 3|x|\} \cup \{(x, y) \in \mathbf{R}^2 \mid |y| < |x|/3\}$$

is plus-open, but it is not open because it contains no ε-ball centered at $(0, 0)$; see Figure 2. Thus, the plus topology is strictly finer than the standard topology.

As evidence that the plus topology is the right topology for studying concepts involving limits with respect to the independent variables separately, we offer the following theorem. The theorem concerns separately continuous functions, where

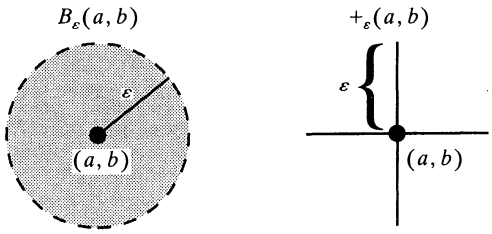


Figure 1. The sets $B_\varepsilon(a,b)$ and $+_\varepsilon(a,b)$.

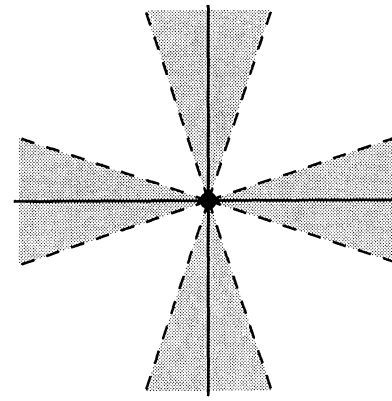


Figure 2. The set A.

a function f with domain $\mathbf{R}^2$ is called *separately continuous* if for every $b \in \mathbf{R}$, the function $f(x,b)$ is a continuous function of x, and for every $a \in \mathbf{R}$, the function $f(a,y)$ is a continuous function of y.

Theorem 1. *For every topological space Y and every function f: $\mathbf{R}^2 \to Y$, f is separately continuous if and only if it is continuous with respect to the plus topology on $\mathbf{R}^2$. Furthermore, the plus topology is the only topology for which this is true.*

Proof: Suppose that f is separately continuous, and let $V \subseteq Y$ be open. Suppose $(a,b) \in f^{-1}(V)$. Then $f(a,b) \in V$, so since the function $f(x,b)$ is continuous, there is some $\varepsilon_1 > 0$ such that if $|x-a| < \varepsilon_1$ then $f(x,b) \in V$. Similarly, there is some $\varepsilon_2 > 0$ such that if $|y-b| < \varepsilon_2$ then $f(a,y) \in V$. Clearly $+_\varepsilon(a,b) \subseteq f^{-1}(V)$, where $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$. Thus $f^{-1}(V)$ is plus-open, so f is continuous with respect to the plus topology.

Now suppose that f is continuous with respect to the plus topology. Suppose that $(a,b) \in \mathbf{R}^2$, and let V be any neighborhood of $f(a,b)$ in Y. Then $(a,b) \in f^{-1}(V)$ and $f^{-1}(V)$ is plus-open, so there is some $\varepsilon > 0$ such that $+_\varepsilon(a,b) \subseteq f^{-1}(V)$. It follows that if $|x-a| < \varepsilon$ then $f(x,b) \in V$, and if $|y-b| < \varepsilon$ then $f(a,y) \in V$. Since V was arbitrary, this shows that the function $f(x,b)$ is continuous at $x = a$ and the function $f(a,y)$ is continuous at $y = b$. Thus, f is separately continuous.

Finally, to prove uniqueness, suppose that T is another topology on $\mathbf{R}^2$ with the property stated in the theorem. Let Y be $\mathbf{R}^2$ with the plus topology, and let f: $\mathbf{R}^2 \to Y$ be the identity function. The f is clearly continuous with respect to the plus topology on the domain, so by the part of the theorem already proved, f must be separately continuous. Thus, f is continuous with respect to the topology T on the domain. In other words, for every plus-open set U, $U = f^{-1}(U) \in T$, so T is at least as fine as the plus topology. Similar reasoning, with the roles of T and the plus topology reversed, shows that the plus topology is at least as fine as T, so T must be the plus topology. ∎

The plus topology is actually a special case of a kind of product topology that has appeared occasionally in the topology literature; see [2] and [3]. There are also

related topologies on $\mathbf{R}^2$ that can be used to study continuity and directional derivatives in directions other than the directions of the coordinate axes. However, in this paper we restrict our attention to the plus topology on $\mathbf{R}^2$.

Corresponding to the fact that all differentiable functions are continuous, we have the following corollary of Theorem 1:

Corollary 2. *Suppose* f: $\mathbf{R}^2 \to \mathbf{R}$. *If the partial derivatives* f_x *and* f_y *are defined everywhere, then* f *is continuous with respect to the plus topology on the domain* $\mathbf{R}^2$.

Proof: If f_x and f_y are defined everywhere then f must be separately continuous, so the conclusion follows from Theorem 1. ■

Since partial derivatives are defined using limits with respect to the independent variables separately, the first partial derivatives of a function f at a point (a, b) can be computed from the values of f at all points in any ε-plus centered at (a, b). Applying this fact at every point in a plus-open set proves our next theorem.

Theorem 3. *Suppose that* f, g: $\mathbf{R}^2 \to \mathbf{R}$, U *is a plus-open set, and for all* $(a, b) \in U$, $f(a, b) = g(a, b)$. *Then for all* $(a, b) \in U$, $f_x(a, b) = g_x(a, b)$ *and* $f_y(a, b) = g_y(a, b)$, *where each equation should be interpreted as meaning that either both partial derivatives are undefined, or both are defined and they are equal.*

Mixed higher order partial derivatives of a function f at a point (a, b) cannot be computed from the values of f on an ε-plus centered at (a, b). However, applying Theorem 3 repeatedly leads to the following corollary:

Corollary 4. *Suppose that* f, g: $\mathbf{R}^2 \to \mathbf{R}$, U *is a plus-open set, and for all* $(a, b) \in U$, $f(a, b) = g(a, b)$. *Then all partial derivatives (including all mixed partials) of* f *and* g *agree at all points in* U.

For example, consider the following two functions:

$$f(x, y) = \begin{cases} x^2 + y^2 & \text{if } (x, y) \in A \\ -1 & \text{if } (x, y) \notin A \end{cases} \qquad g(x, y) = x^2 + y^2, \tag{1}$$

where A is the plus-open set in Figure 2; see Figures 3 and 4. These functions agree at all points in A, so by Corollary 4 their partial derivatives of all orders also agree at all points in A. In particular, all partial derivatives of f and g agree at $(0, 0)$. We might say that the partial derivatives at $(0, 0)$ look at points only in a plus-open neighborhood of $(0, 0)$, and therefore they don't see the difference between f and g. But the local (in the sense of the standard topology) behavior of these functions is quite different near $(0, 0)$. For example, g is differentiable at $(0, 0)$, and f is not even continuous there. This illustrates the point that partial derivatives of a function do not give information about its local behavior.

This example also makes it clear that it is impossible to tell whether or not a function is differentiable at a particular point by examining its partial derivatives (of any order) at that point. The test for differentiability given in most multivariable calculus books says that a function is differentiable at a point if the first partial derivatives are not only defined but also *continuous* at that point. In fact, examination of the proof shows that it suffices to assume that only *one* of the partial derivatives is continuous, but this example shows why one cannot drop the continuity requirement completely.

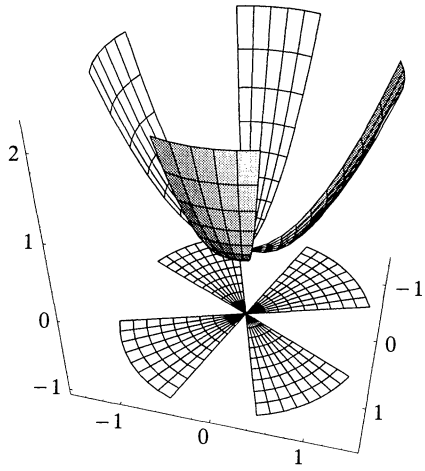


Figure 3. $z = f(x, y)$.

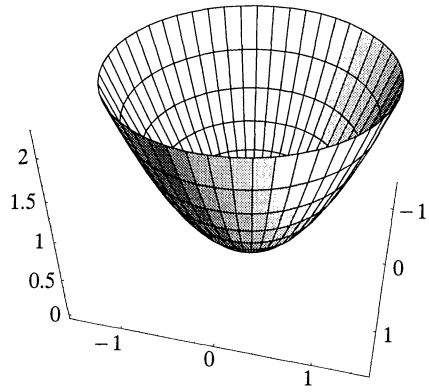


Figure 4. $z = g(x, y)$.

Here is another well-known theorem from multivariable calculus; see [**5**, p. 212]:

Theorem 5. (Second Derivative Test for Local Extrema) *Suppose that* $f(x, y)$ *is differentiable in a neighborhood of* (a, b), $f_x(a, b) = f_y(a, b) = 0$, *and* f_x *and* f_y *are differentiable at* (a, b). *Let* $D = f_{xx}(a, b)f_{yy}(a, b) - [f_{xy}(a, b)]^2$. *Then*:

1. *If* $D > 0$ *and* $f_{xx}(a, b) > 0$ *then* f *has a local minimum at* (a, b).
2. *If* $D > 0$ *and* $f_{xx}(a, b) < 0$ *then* f *has a local maximum at* (a, b).
3. *If* $D < 0$ *then* f *does not have a local extremum at* (a, b).

Once again, the plus topology can be helpful in constructing and understanding examples that illustrate why the hypotheses are needed. It is easy to check that the Second Derivative Test correctly determines that the function g in (1) has a local minimum at $(0, 0)$. Since the partial derivatives of f and g in (1) agree at $(0, 0)$, the test gives the same answer for f, even though f does not have a local minimum at $(0, 0)$. Of course, f does not satisfy the first hypothesis of Theorem 5, since it is not differentiable in a neighborhood of $(0, 0)$. But it is not hard to modify f to make it differentiable everywhere, and still have the Second Derivative Test fail. We simply need a surface that is the same as the graph of g on a plus-open neighborhood of $(0, 0)$, but is concave downward outside of that neighborhood. A natural choice would be a surface given in polar coordinates by an equation of the form $z = c(\theta)r^2$, where $c(\theta)$ is 1 when θ is close to an integer multiple of $\pi/2$ and $c(\theta)$ changes smoothly to a negative value when θ is an odd multiple of $\pi/4$. For example, we might let c be a function that is periodic with period $\pi/2$ and define $c(\theta)$ for θ between 0 and $\pi/2$ as follows:

$$c(\theta) = \begin{cases} 1 & \text{if } 0 \le \theta \le \pi/8 \\ 1 - \exp\left[6 + \dfrac{1}{\theta - 3\pi/8} - \dfrac{1}{\theta - \pi/8}\right] & \text{if } \pi/8 < \theta < 3\pi/8 \\ 1 & \text{if } 3\pi/8 \le \theta \le \pi/2. \end{cases}$$

The graph of c is shown in Figure 5, and the surface $z = c(\theta)r^2$ is shown in Figure 6. This surface is the graph of a function $h(x, y)$ that is infinitely differentiable at all points other than the origin, since $c(\theta)$ is infinitely differentiable, and

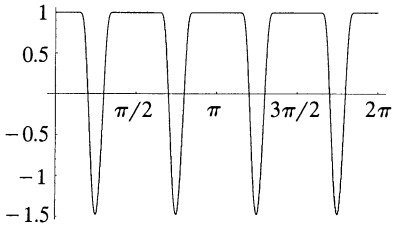


Figure 5. $y = c(\theta)$.

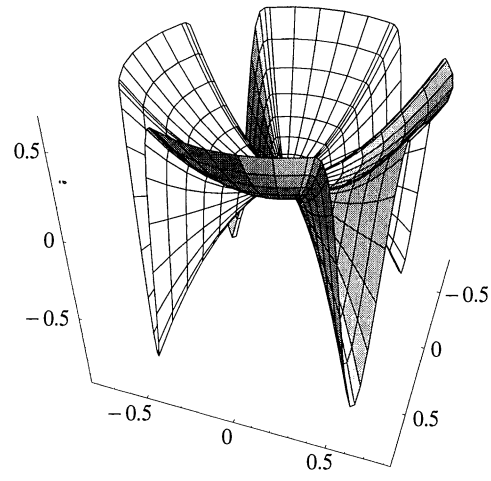


Figure 6. $z = c(\theta)r^2$.

θ and r are infinitely differentiable functions of x and y in a neighborhood of any point except the origin. Furthermore, since $c(\theta)$ is bounded, there are constants a and b such that $a(x^2 + y^2) \leq h(x, y) \leq b(x^2 + y^2)$, from which it follows that h is differentiable at $(0, 0)$. And finally, since h agrees with g on a plus-open neighborhood of $(0, 0)$, all partial derivatives of h are defined at $(0, 0)$ and are the same as the partial derivatives of g. Therefore the Second Derivative Test incorrectly indicates a local minimum for h at $(0, 0)$. The only hypothesis of Theorem 5 that we have not checked is the differentiability of the partial derivatives at $(0, 0)$, so this hypothesis must fail for h, and it cannot be dropped from the theorem. The reader might enjoy checking that $h_x(x, y) = 2xc(\theta) - yc'(\theta)$ and $h_y(x, y) = 2yc(\theta) + xc'(\theta)$. Using the fact that $c(\theta)$ and $c'(\theta)$ are bounded but not constant, it can be shown that the first partial derivatives are continuous but not differentiable at $(0, 0)$. One can get an example where the Second Derivative Test incorrectly indicates that a function does not have a local extremum by adding $z = (1 - c(\theta))r^2$ to an appropriately chosen surface with a saddle at $(0, 0)$, such as $z = x^2 + y^2 + (2 + \varepsilon)xy$, for sufficiently small positive ε. Similar examples can be found in [**4**].

All of our examples so far have been based on the plus-open set A, but there are many more exotic plus-open sets. For example, let $\{B_1, B_2, B_3, \ldots\}$ be a countable basis for the standard topology on $\mathbf{R}^2$. Inductively choose, for each positive integer n, a point $(x_n, y_n) \in B_n$ such that for all $m < n$, $x_n \neq x_m$ and $y_n \neq y_m$. Let $F = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\}$. Since F contains a point from every basic open set, $\mathbf{R}^2 \setminus F$ has empty interior in the standard topology. However, we claim that $\mathbf{R}^2 \setminus F$ is plus-open. To see why, suppose $(a, b) \in \mathbf{R}^2 \setminus F$. Then since there is at most one point in F with y-coordinate b, it is easy to find an $\varepsilon_1 > 0$ such that if $|x - a| < \varepsilon_1$ then $(x, b) \notin F$. Similarly, we can find an $\varepsilon_2 > 0$ such that if $|y - b| < \varepsilon_2$ then $(a, y) \notin F$. Thus $+_\varepsilon(a, b) \subseteq \mathbf{R}^2 \setminus F$, where $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$.

Unusual plus-open sets can lead to unusual examples in multivariable calculus. For example, define $j\colon \mathbf{R}^2 \to \mathbf{R}$ as follows:

$$j(x, y) = \begin{cases} 1 & \text{if } (x, y) \in F \\ 0 & \text{if } (x, y) \notin F. \end{cases} \tag{2}$$

Then j agrees with the constant function $c(x, y) = 0$ on the plus-open set $\mathbf{R}^2 \setminus F$, and therefore by Corollary 4 all partial derivatives of j are defined and equal to 0 everywhere except for the countably many points in F. But since both F and $\mathbf{R}^2 \setminus F$ are dense in the plane in the standard topology, j is discontinuous everywhere.

Can the countable set of exceptional points in this example be avoided? Can a function have partial derivatives defined everywhere but be discontinuous everywhere? The answer is no, but to see why we need a fact about closures in the plus topology. For a set $X \subseteq \mathbf{R}^2$, we write $\mathrm{cl}(X)$ for the closure of X in the standard topology, and $\mathrm{cl}_+(X)$ for the closure of X in the plus topology. Note that $\mathrm{cl}_+(X) \subseteq \mathrm{cl}(X)$, since the plus topology is finer than the standard topology. For $X \subseteq \mathbf{R}$ we also write $\mathrm{cl}(X)$ for the closure of X in the standard topology on $\mathbf{R}$.

Our example $\mathbf{R}^2 \setminus F$ shows that a nonempty plus-open set can have empty interior in the standard topology. However, this cannot be true of the closure in the plus topology of a nonempty plus-open set. In fact, we have the following slightly stronger theorem, which implies that $\mathbf{R}^2$ with the plus topology is a Baire space:

Theorem 6. *Suppose U is a nonempty plus-open set, and $U = \bigcup_{n \in \mathbf{Z}^+} U_n$. Then for some n, $\mathrm{cl}_+(U_n)$ has nonempty interior in the standard topology.*

Proof: Let $(a, b) \in U$, and choose $\varepsilon > 0$ such that $+_\varepsilon(a, b) \subseteq U$. For each $x \in (a - \varepsilon, a + \varepsilon)$ and $n \in \mathbf{Z}^+$, let $Y_n^x = \{y | (x, y) \in U_n\}$, and let $Y^x = \bigcup_{n \in \mathbf{Z}^+} Y_n^x = \{y | (x, y) \in U\}$. Since $(x, b) \in +_\varepsilon(a, b) \subseteq U$ and U is plus-open, Y^x must contain an interval. Thus, by the Baire Category Theorem, there is some positive integer n_x such that $\mathrm{cl}(Y_{n_x}^x)$ contains an interval. Choose rational numbers p_x and q_x such that $p_x < q_x$ and $(p_x, q_x) \subseteq \mathrm{cl}(Y_{n_x}^x)$.

For each positive integer n and rational interval (p, q), let $X_{n,p,q} = \{x \in (a - \varepsilon, a + \varepsilon) | n_x = n,\ p_x = p, \text{ and } q_x = q\}$. Since there are only countably many possible values for n, p, and q, another application of the Baire Category Theorem shows that there must be some n, p, and q such that $\mathrm{cl}(X_{n,p,q})$ contains an interval. Choose $c < d$ such that $(c, d) \subseteq \mathrm{cl}(X_{n,p,q})$. For each $x \in X_{n,p,q}$, $(p, q) \subseteq \mathrm{cl}(Y_n^x)$, and it is not hard to see that therefore $X_{n,p,q} \times (p, q) \subseteq \mathrm{cl}_+(U_n)$. Similarly, since $(c, d) \subseteq \mathrm{cl}(X_{n,p,q})$, it follows that $(c, d) \times (p, q) \subseteq \mathrm{cl}_+(U_n)$, as required. ∎

Using Theorem 6, we can prove the following theorem of Baire; see [**1**] and [**6**]:

Theorem 7. (Baire) *Suppose f: $\mathbf{R}^2 \to \mathbf{R}$ and suppose f_x and f_y are defined at all points in $\mathbf{R}^2$. Then there is a dense set of points at which f is differentiable.*

Proof: For $h \neq 0$ define functions m_h and n_h as follows:

$$m_h(x, y) = \frac{f(x + h, y) - f(x, y)}{h}, \qquad n_h(x, y) = \frac{f(x, y + h) - f(x, y)}{h}.$$

Note that m_h and n_h are separately continuous, since f is. Of course $\lim_{h \to 0} m_h(x, y) = f_x(x, y)$ and $\lim_{h \to 0} n_h(x, y) = f_y(x, y)$.

We claim first that if V is any nonempty open set and $\varepsilon > 0$ then there is a nonempty open set W such that $\mathrm{cl}(W) \subseteq V$ and for all $(u, v), (x, y) \in W$, $|f_x(u, v) - f_x(x, y)| < \varepsilon$ and $|f_y(u, v) - f_y(x, y)| < \varepsilon$. To prove the claim, first choose a nonempty open set X such that $\mathrm{cl}(X) \subseteq V$. Now for each positive integer

n and rational numbers p and q, let

$$U_{n,p,q} = \{(x,y) \in X | \text{ for all } h, \text{ if } 0 < |h| < 1/n \text{ then } |m_h(x,y) - p| < \varepsilon/3 \text{ and } |n_h(x,y) - q| < \varepsilon/3\}.$$

Clearly $\bigcup\{U_{n,p,q} | n \in \mathbf{Z}^+ \text{ and } p, q \in \mathbf{Q}\} = X$, so by Theorem 6 we can choose $n \in \mathbf{Z}^+$ and $p, q \in \mathbf{Q}$ such that $\mathrm{cl}_+(U_{n,p,q})$ has nonempty interior in the standard topology. Let W be the interior of $\mathrm{cl}_+(U_{n,p,q})$. Then $\mathrm{cl}(W) \subseteq \mathrm{cl}(X) \subseteq V$, and using the fact that m_h and n_h are separately continuous, it is not hard to see that

$$W \subseteq \mathrm{cl}_+(U_{n,p,q}) \subseteq \{(x,y) \in \mathbf{R}^2 | \text{for all } h, \text{ if } 0 < |h| < 1/n \text{ then } |m_h(x,y) - p| \leq \varepsilon/3 \text{ and } |n_h(x,y) - q| \leq \varepsilon/3\}.$$

It follows that for all $(x,y) \in W$, $|f_x(x,y) - p| \leq \varepsilon/3$ and $|f_y(x,y) - q| \leq \varepsilon/3$, and therefore for all $(u,v),(x,y) \in W$, $|f_x(u,v) - f_x(x,y)| \leq 2\varepsilon/3 < \varepsilon$ and $|f_y(u,v) - f_y(x,y)| \leq 2\varepsilon/3 < \varepsilon$, as required.

Now let V_0 be any nonempty bounded open set. To prove the theorem, we must find a point in V_0 at which f is differentiable. By the claim, let V_1 be a nonempty open set such that $\mathrm{cl}(V_1) \subseteq V_0$ and for all $(u,v),(x,y) \in V_1$, $|f_x(u,v) - f_x(x,y)| < 1$ and $|f_y(u,v) - f_y(x,y)| < 1$. Applying the claim again, let V_2 be a nonempty open set such that $\mathrm{cl}(V_2) \subseteq V_1$ and for all $(u,v),(x,y) \in V_2$, $|f_x(u,v) - f_x(x,y)| < 1/2$ and $|f_y(u,v) - f_y(x,y)| < 1/2$. In general, given V_n we choose a nonempty open set V_{n+1} such that $\mathrm{cl}(V_{n+1}) \subseteq V_n$ and for all $(u,v),(x,y) \in V_{n+1}$, $|f_x(u,v) - f_x(x,y)| < 1/(n+1)$ and $|f_y(u,v) - f_y(x,y)| < 1/(n+1)$.

Let $(a,b) \in \bigcap_{n \in Z^+} V_n$. Then for every positive integer n, $(a,b) \in V_n$, and for every $(x,y) \in V_n$, $|f_x(a,b) - f_x(x,y)| < 1/n$ and $|f_y(a,b) - f_y(x,y)| < 1/n$. It follows that f_x and f_y are continuous at (a,b), and therefore f is differentiable at (a,b), as required. ∎

Returning to our function j in (2), we can now see why the exceptional points cannot be avoided. If the partial derivatives of a function are defined everywhere then, by Theorem 7, it must be not only continuous but also differentiable at a dense set of points. Our function j shows that the hypotheses of Theorem 7 cannot be weakened to allow a countable set of exceptional points.

We close by mentioning two unusual properties of the plus topology that distinguish it from the standard topology on $\mathbf{R}^2$. The first follows almost immediately from Theorem 6:

Theorem 8. *The plus topology is not regular.*

Proof: We have already seen that $\mathbf{R}^2 \setminus F$ is plus-open, so F is plus-closed. Let (a,b) be any point not in F. We claim that (a,b) and F cannot be separated by plus-open sets. To see why, suppose that U and V are disjoint plus-open sets with $(a,b) \in U$ and $F \subseteq V$. Then $\mathrm{cl}_+(U)$ has empty interior in the standard topology, contradicting Theorem 6. ∎

The second unusual property of the plus topology is that it is not second countable, or even first countable. In fact, it is surprisingly difficult to find a natural basis for the plus topology. Note that the sets $+_\varepsilon(a,b)$ are not plus-open, and therefore cannot be used as basis sets. It turns out that for any point $(a,b) \in \mathbf{R}^2$, any local basis at (a,b) for the plus topology must have $2^{2^{\aleph_0}}$ elements. This follows from more general results in [2].

ACKNOWLEDGMENT. I am grateful to Joan Hart, Ken Kunen, and Norton Starr for helpful comments on earlier drafts of this paper.

REFERENCES

1. R. Baire, Sur les fonctions des variables réelles, *Ann. Mat. Pura Appl.* **3** (1899) 1–122.
2. J. Hart and K. Kunen, On the regularity of the topology of separate continuity, to appear in *Topology Appl.*
3. C. J. Knight, W. Moran, and J. S. Pym, The topologies of separate continuity, I, *Proc. Camb. Phil. Soc.* **68** (1970) 663–671.
4. A. A. Struthers, Counterexamples to a weakened version of the two-variable second derivative test, *College Math. J.* **28** (1997) 383–385.
5. A. E. Taylor and W. R. Mann, Advanced Calculus, 3rd ed., Wiley, New York, 1983.
6. E. B. van Vleck, A proof of some theorems on pointwise discontinuous functions, *Trans. Amer. Math. Soc.* **8** (1907) 189–204.

DAN VELLEMAN received his B.A. from Dartmouth College in 1976 and his Ph.D. from the University of Wisconsin in 1980. He taught at the University of Texas and the University of Toronto before joining the faculty of Amherst College in 1983. He is the author of the book *How to Prove It* and the accompanying software ProofDesigner (which can be downloaded from http://www.cs.amherst.edu/~djv), and a coauthor, together with Joe Konhauser and Stan Wagon, of the book *Which Way Did the Bicycle Go?* He is interested in logic, the philosophy of mathematics, and the foundations of quantum mechanics.
Department of Mathematics and Computer Science, Amherst College, Amherst, MA 01002
djvelleman@amherst.edu

The Forced Damped Pendulum: Chaos, Complication and Control

John H. Hubbard

We show that a "simple" differential equation modeling a garden-variety damped forced pendulum can exhibit extraordinarily complicated and unstable behavior. While instability and control might at first glance appear contradictory, we can use the pendulum's instability to control it. Such results are vital in robotics: the forced pendulum is a basic subsystem of any robot.

Most of the mathematical methods used in this paper were initially developed in celestial mechanics, largely by Poincaré. The literature of the field tends to be quite advanced indeed (see [**1**] and [**11**]); one object of this paper is to show that computer programs, properly used, can make these advanced topics transparent. All the computer-generated pictures in this paper were produced by the programs *Planar Systems* and *Planar Iterations* [**6**], both written by Ben Hinkle (now at *Maple*).

1. SOME PARALLELS IN CELESTIAL MECHANICS. When I was a graduate student, I was amazed by the results of Alekseev concerning a system formed by three bodies obeying Newton's law of gravitation; see [**1**] and [**11**]. As shown in Figure 1, two massive bodies of equal mass move in a plane P on ellipses

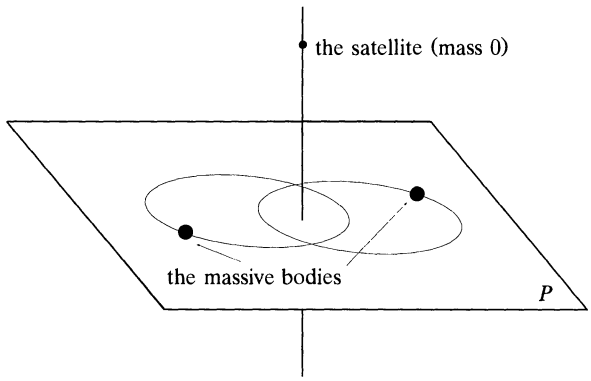


Figure 1. Alekseev's three-body system.

symmetric around a common focus F, and the third body, the *satellite*, of mass zero, moves on the line L perpendicular to P through F. Once this satellite is launched, its motions are determined uniquely by the gravitational pull of the two massive bodies.

The system has a natural unit of time, the "year"—the time it takes the massive bodies to complete a revolution. Choose a time zero, so that it makes sense to speak of the 0th, 1st, ..., nth year. Also let x denote the position on the line L, with $x = 0$ corresponding to F.

Alekseev proved that there then exists a number N, which depends on the eccentricity of the orbits of the large bodies, such that given *any* sequence $n_1, n_2, \ldots$ of integers at least N, there exists a set of initial conditions that results in the satellite returning to cross the plane P exactly in the n_1th year, the $(n_1 + n_2)$th year, etc. In other words, given a specified sequence of years with gaps at least N, it is possible to choose an instant t_0 and a speed $v = x'(t_0)$ so that if the satellite is kicked off at that moment with that speed, it crosses the plane during the desired years: first during the n_1th year, then n_2 years later, and so on. You can set up the satellite to return in *any* sequence of years you like, so long as the returns are spaced at least N apart.

In particular, there exist unbounded orbits in which the satellite travels arbitrarily far away but always returns, for example the orbit corresponding to the sequence of gaps between crossings $N, N + 1, N + 2, N + 3, \ldots)$ as well as infinitely many different periodic orbits (for instance $N, N + 12, N + 17, N, N + 12, N + 17, \ldots)$.

Actually, Alekseev claimed the result only when the eccentricity is "sufficiently small." He needed to know that his system satisfied some requirements (basically, that a "horseshoe" should be present), and he could verify this only by a perturbation calculation near an explicitly integrable system. Horseshoes are discussed in Section 8.

The pendulum model we explore here exhibits a similar sort of behavior: we can make our pendulum go through any specified sequence of gyrations by correctly choosing the initial conditions. More precisely, by appropriately choosing the position and the velocity of the pendulum at time 0, we can specify whether during each time period (the time period of the forcing term, in our case, 2π) the pendulum goes through the bottom position once clockwise, once counterclockwise, or not at all. For example, we could specify that in each of the first six periods it could go through the bottom position once clockwise, in each of the next three periods it could go through the bottom position once counterclockwise, and in the tenth period oscillate around an upright position.... All imaginable sequences are possible: once the correct set of initial conditions is chosen, the differential equation governing the system automatically enforces the desired behavior.

2. DIFFERENTIAL EQUATIONS AND PENDULUMS. There is only one law in mechanics: $F = ma$ (force equals mass times acceleration). Thus the motion of a pendulum of length l, with a bob of mass m in a constant gravitational field of force g, with friction proportional to the velocity, and forcing $f(t)$ (Figure 2) is modeled by the differential equation

$$\underbrace{f(t) - \gamma l x' - mg \sin(x)}_{\text{force}} = \underbrace{m}_{\text{mass}} \times \underbrace{lx''}_{\text{acceleration}} .$$

The friction term $\gamma l x'$ is a fairly good approximation to reality when the friction is due to air, and the speed of the bob is much less than the speed of sound. The term $mg \sin(x)$ is the force exerted by gravity; the weight of the body is mg, but only the component in the direction of motion contributes to the equation. The forcing $f(t)$ can be created by a current proportional to $f(t)$ through the axis of the pendulum, if the bob is a bar magnet perpendicular to the axis. In realistic situations (e.g., robot arms), this is the way forcing is really produced.

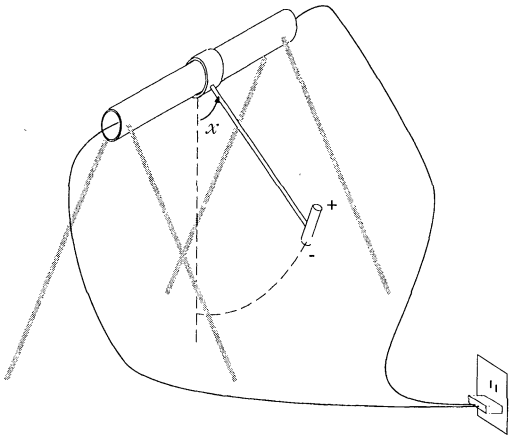


Figure 2. A pendulum being driven by alternating current.

We explore the behavior of a pendulum whose motions are described by the particular differential equation

$$\cos(t) - 0.1x' - \sin(x) = x'',$$

in which both mass m and length l equal 1.

My starting point was the observation by Borelli and Coleman [**3**] that numerical solutions of this equation are very sensitive to the integration method, step-length, etc., near the initial condition $(x(0), x'(0)) = (0, 2)$. That is, we start with a pendulum hanging down, and hit it with a mallet to give it velocity near 2. This paper is my attempt to understand this instability. The behavior I describe holds not just for the parameters $m, \gamma, l, g, f(t)$ given; they could be varied in a certain range, which I don't know in any detail, but which is large enough so that it would not be difficult to build a real system that behaves like the one described here.

3. A FIRST ATTEMPT TO UNDERSTAND THE MOTIONS OF THE PENDULUM. The most obvious thing to ask a computer is: what do the motions of the pendulum look like? The following picture shows the motion resulting from 15 different sets of initial conditions. Each graph starts with the position $x(0) = 0$; the initial velocities are evenly spaced between 1.85 and 2.1. The graphs are plotted for $-1 < t < 200$ and $-25 < x < 25$. A word of caution: the overall features of Figure 3 are correct, but the details—exactly which equilibrium each initial condition leads to—might well be wrong. The exponential growth of errors is discussed in Section 11.

A careful look at the picture suggests that there exists a stable periodic motion $S(t)$ of the pendulum, which you see in the picture many times; of course, $S(t) + 2k\pi$ is another description of the same motion for any integer k; the letter S stands for "stable." You will see five different levels of this stable periodic motion: one on the horizontal axis, three above, and one below. The first stable motion above the horizontal axis represents motions that go "over the top" once counterclockwise before settling down, like a child's swing going over the bar. The next layer up represents motions that go over the top twice counterclockwise before settling down, while the layer below the horizontal axis represents motions that go over the top once clockwise before settling down.

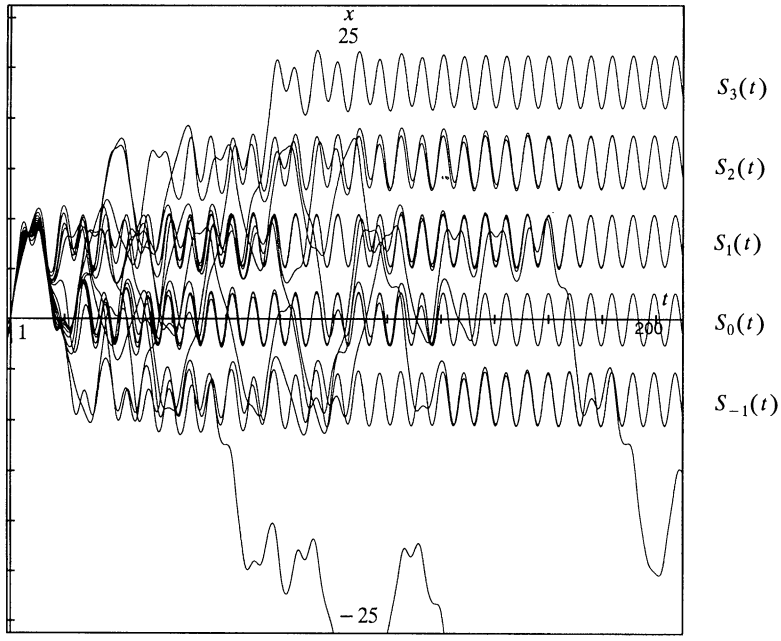


Figure 3. Fifteen solutions to the differential equation $\cos(t) - 0.1x' - \sin(x) = x''$.

Some motions rapidly settle down to this oscillation, others go through a complicated path before doing so, and yet others do not approach the periodic motion in this amount of time. These appear to be rare, and one might guess that given more time, almost all solutions do settle down. (One that does not is shown in [**13**, p. 228]; the existence of uncountably many others is proved in Theorem 3.)

An obvious question is: what stable oscillation—what attracting periodic solution—can a motion approach? This seems impossible to understand without another program.

4. THE SCANNING PICTURE. We now look at the whole family of initial conditions: position represented by the horizontal axis, velocity by the vertical axis. We ask the computer to color initial conditions according to the stable oscillation the corresponding solution approaches (if any). This set of initial conditions is called the *basin* of the corresponding *sink*; it is an open subset of $\mathbb{R}^2$.

This is best done as follows. First, find the initial values $S_0(0), S_0'(0)$ for one of the attracting periodic solutions, say the one with $-2\pi < S_0(0) < 0$. We call the motion immediately above it S_1, and the one above that S_2; we have $S_k(t) = S_0(t) + 2k\pi$. Next, find a number $r > 0$ such that if

$$|x(0) - S_0(0)|^2 + |x'(0) - S_0'(0)|^2 < r^2,$$

then the motion $x(t)$ is definitely attracted to S_0. That is, any set of initial values inside the circle of radius r and centered at $(S_0(0), S_0'(0))$, gets arbitrarily close to the solution S_0 (in fact, does so exponentially fast). We rely on computer calculations to determine this, but it would not be hard to provide a rigorous mathematical justification. We are not particularly interested in the points inside that circle; we are just establishing how we know that a motion is attracted to a particular attracting solution: it is attracted to it if it ever enters the circle of radius r around

the solution. In our case, we have

$$(S_0(0), S_0'(0)) \approx (-2.0463, .3927) \quad \text{and we can take} \quad r = 0.1.$$

Now we solve the differential equation starting at every point of some grid (in our case, a 600×400 grid—240,000 points!), and sample the solution at times $2\pi, 4\pi, \ldots$: this is a substantial computation, taking about two hours even on a fairly fast Macintosh (200 MHz).

If for some such motion $w(t)$ and some integer $n > 0$ we have

$$|w(2n\pi) - S_k(0)|^2 + |w'(2n\pi) - S_k'(0)|^2 < r^2,$$

we know that this motion is attracted to S_k. Color the point $(w(0), w'(0))$ in the kth color and solve the differential equation for the next point. If after some number of samplings (in our case 30: we integrated solutions for time $60\pi \approx 185$) the solution never falls within r of an attracting solution, leave the initial point white. We obtain Figures 4 and 5.

5. LAKES OF WADA. The colored sets B_k (called, for obvious reasons, the *basins* of the corresponding attracting motions S_k are immensely complicated.

We show that they form infinitely many Lakes of Wada. Wada was a Japanese mathematician who at the beginning of the 20th century constructed an example of three disjoint, connected open subsets of the unit disc $D \subset \mathbb{R}^2$ such that every point in the boundary of one is in the boundary of the other two [**15**]. This amazed the mathematical community at the time: if you try to draw three (connected, open) lakes in an island, you would probably soon convince yourself that all three can touch at only two points. Actually, it appears that Brouwer discovered this phenomenon earlier [**4**].

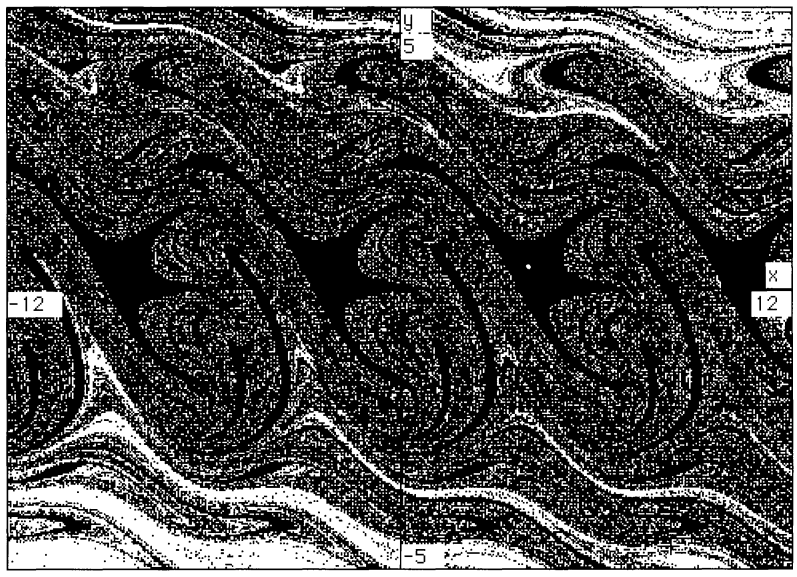


Figure 4. The different colors (hard to appreciate in black and white) represent different basins: which initial conditions are attracted to which sinks. Points colored white may be initial conditions that are never attracted to a sink, but more likely they are attracted to sinks that are off the picture. They could also be attracted to sinks in the picture, but not during the time allowed.

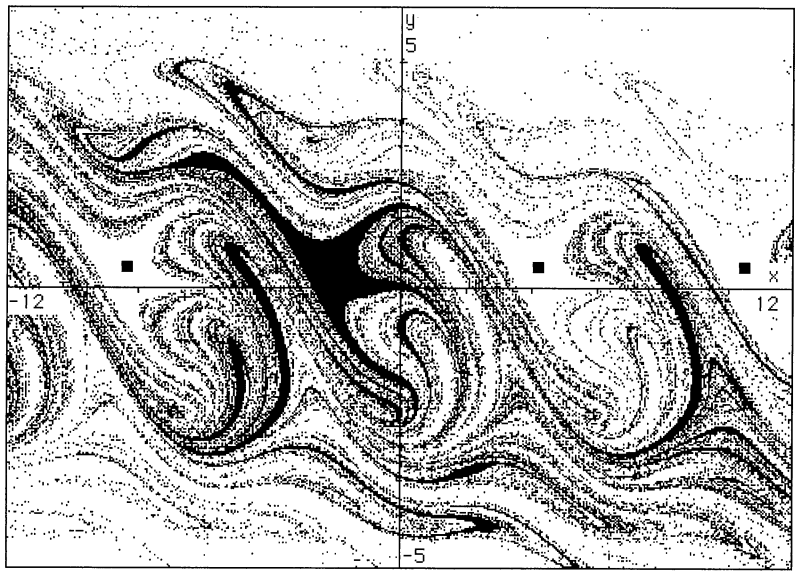


Figure 5. In black and white, the four basins of Figure 4 are hard to distinguish. This figure represents just one basin.

Let me sketch the construction as outlined in [**15**], illustrating the dangers of philanthropy; this is illustrated by Figure 6.

Suppose D is an island cursed with three philanthropists, one of whom wants to bring water to every inhabitant, one tea, and one coffee. At the beginning each has a pond of his own beverage.

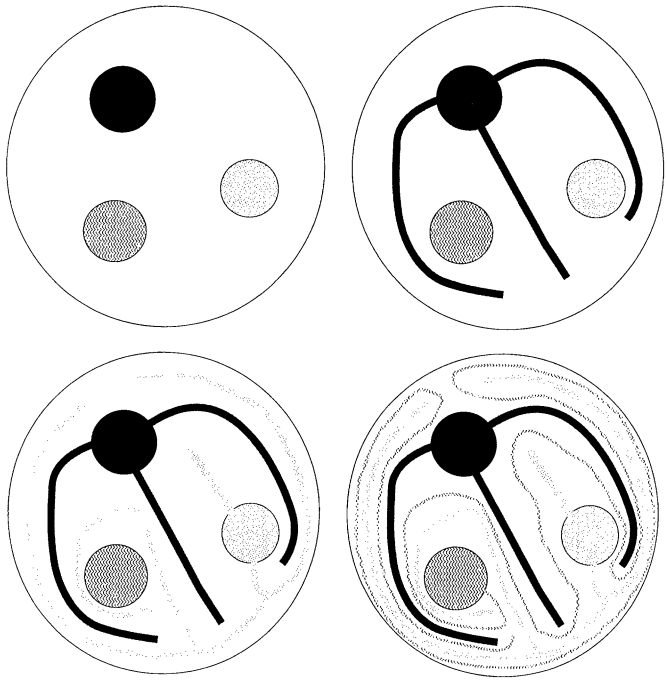


Figure 6. Digging the lakes of Wada.

First, the purveyor of water digs a system of canals emanating from his pond, and bringing water within 100 meters of every inhabitant, never actually touching the surrounding sea or the other ponds, and forming no loops.

Next, the purveyor of coffee builds a system of canals emanating from his pond, bringing coffee to within 10 meters of every inhabitant, again forming no loops. Since the water canals make no loops, they don't cut off any inhabitants from the coffee pond, so this is possible.

Now the purveyor of tea builds his system of canals, bringing tea to within 1 meter of every inhabitant. Next the water purveyor goes back to work, extending his canals (necessarily building narrower ones) to bring water within 10 cm of each inhabitant. And so forth. At the end of this process, the poor inhabitants no longer have any dry land to stand on, but they have water, tea, and coffee as close as they want. What remains of the dry land is in the boundary of all three basins.

Real philanthropists don't seem to behave this way, fortunately. Highway designers, on the other hand...

Theorem 1 shows that our pendulum is creating lakes of Wada.

Theorem 1. *The basins B_k have the Wada property: every point in the boundary of one basin is also in the boundary of all the others.*

This is not quite as strong as the preceding statement about philanthropists, where every bit of dry land was in the boundary of all the basins. For the pendulum, all we can prove is that *if* a point is in the boundary of one basin, it's in the boundary of the others. Presumably there is no other dry land, but we don't know how to prove it. True lakes of Wada have been proven to exist in another setting of dynamical systems [**7**].

The first step in understanding why Theorem 1 is true is to get a grasp on the boundaries of the basins. Most of the material in the next section was developed by Kennedy, Nusse and Yorke; see [**9**] and [**12**]. They saw that the basin of a sink often has *saddle points* on its boundary, and that the *stable separatrices* of these saddle points make up the *accessible boundary* of the basin. We will first define these words.

6. ITERATION, SINKS, SADDLES, SEPARATRICES. Rather than thinking of the differential equation in $\mathbb{R}^3$, I find it much easier to think of the *period mapping* (or *Poincaré mapping*) in the plane

$$P:\mathbb{R}^2 \to \mathbb{R}^2 \quad \text{given by} \quad P:\begin{bmatrix} x(0) \\ x'(0) \end{bmatrix} \mapsto \begin{bmatrix} x(2\pi) \\ x'(2\pi) \end{bmatrix}.$$

This enables me to ignore what motions do between the samples.

There is no real loss if we are interested in long-term behavior: *iterating* m times the mapping P is equivalent to solving the differential equation for time $2m\pi$, sampling the solutions every 2π. But the dynamical objects are now subsets of the plane rather than of space: most people visualize objects in the plane much better than in space. In our case, the planar objects are quite complicated enough.

Seen this way, each point $s_k = (S_k(0), S_k'(0))$ is an attracting fixed point of P, also called a *sink*: $P(s_k) = s_k$ and if a point p is close to s_k (within r of it, for instance), its orbit under P approaches s_k. The basin B_k is exactly the set of points p such that the sequence $p, P(p), P^2(p), \ldots$ approaches s_k.

Sinks can also be periodic of period $m > 1$. Such sinks are points p such that $P^m(p) = p$, and such that if a point p_1 is sufficiently close to p, the sequence, $p_1, P^m(p_1), P^{2m}(p_1), \ldots$ tends to p. That is, the solution of the differential equation with $(x(0), x'(0)) = p$ is an attracting periodic solution of period $2m\pi$. Our mapping P appears not to have any such points (for these values of the parameters), although proving that it has none may well be an unsolvable problem. But there are infinitely many periodic saddles, as is proved by Theorem 3. And there are infinitely many more whose existence is not guaranteed by that theorem.

Like a sink, a *saddle point* for P corresponds to a periodic solution of the original differential equation, but while sinks are associated with stable equilibria, saddles are associated with unstable equilibria. A periodic solution $(x(t), x'(t))$ of the differential equation gives a saddle $(x(0), x'(0))$ of the period mapping P if there is a surface made up of solutions of the differential equation that tend to the attracting periodic solution as time tends to $+\infty$, and another surface of solutions that tend to the attracting periodic solution as $t \to -\infty$, i.e., as one travels backwards in time.

An example of a saddle point is the upwards (unstable) equilibrium for an unforced damped pendulum. Almost all solutions are captured by a stable equilibrium. But exceptional solutions exist that take an infinite amount of time to approach the vertical, and other solutions take an infinite amount to fall away from the vertical: these solutions make up two surfaces that intersect along the constant solution corresponding to the unstable equilibrium. The surface of solutions that tend to the vertical in forward time is the *stable separatrix*, while the surface of solutions tending to the vertical in backwards time is the *unstable separatrix*. The intersection of these surfaces with a Poincaré plane (i.e., the plane $t = 0$) forms two curves, also referred to as separatrices. Think of the separatrices as watersheds: for our unforced pendulum, they separate the initial conditions that go over the top one more time from those that don't make it.

Mappings $\mathbb{R}^2 \to \mathbb{R}^2$ (which might be the period mapping of a time-periodic differential equation in $\mathbb{R}^2$, as in our case) usually also have *sources*: fixed or periodic points that repel all nearby orbits. The period mapping P for our pendulum has no sources because P contracts areas by $e^{-2\pi/10} \approx 0.53$, due to the damping [**8**, vol. 2, chap. 8]. No mapping can simultaneously contract areas and map some region to a strictly larger region, as would have to happen near a source. Of course, P^{-1} has sources wherever P has sinks.

7. SADDLES IN THE BOUNDARY OF B_K. The computer finds four saddles $p_{k,1}, \ldots, p_{k,4}$ in the boundary of each basin. These saddles form two cycles of period 2 (i.e., the solutions of the differential equation with initial values at these saddles have period 4π). The boundary of the basin appears to be made of their stable separatrices, as drawn in Figure 7. We will call these separatrices $\sigma^+(p_{k,i})$: these are the watersheds that separate the solutions falling into the basin from those that don't.

In fact, the preceding statement is not true: the boundaries of the basins are not just the separatrices; they are much more complicated than that. The complication stems from the fact that all points of the boundary are limits of *sequences in the basin*, but not all such points are limits of *paths*. Consider Wada's construction: some points of the boundary of the water are on the edge of some water stream, but most are not. For one thing, points on the edge of a coffee stream are not on the edge of a water stream, even though they are in the boundary of the water: there are water streams arbitrarily close, but tea streams even closer, etc. Such

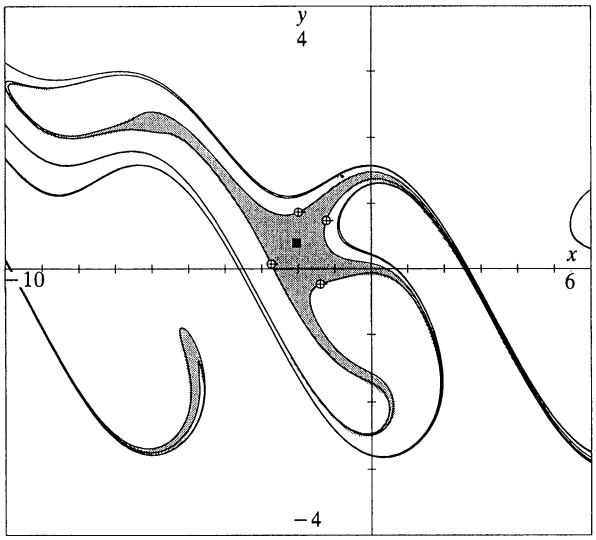


Figure 7. The stable separatrices of the saddles of period 2 in the boundary of a basin provide an outline drawing of the basin. Thus this picture is more or less the same as Figure 5, but the stable manifolds would need to be continued for a very long time to get as much resolution as figure 5 provides.

points are *inaccessible* by water: you can reach out to them over other streams, with an arbitrarily small motion, but you cannot reach them in a boat. Most points of the common boundary (the *separator)* are not accessible from the water, coffee, or tea.

Our basins are similar to those of the Wada example. Each includes a central "pond" with four canals leading off from it, which dwindle to become infinitely narrow streams, intermingled with streams belonging to other basins.

In our case, the *inward pointing* unstable separatrix at each of the four saddles is attracted to the sink, as shown in Figure 8, and provides a path from the sink to

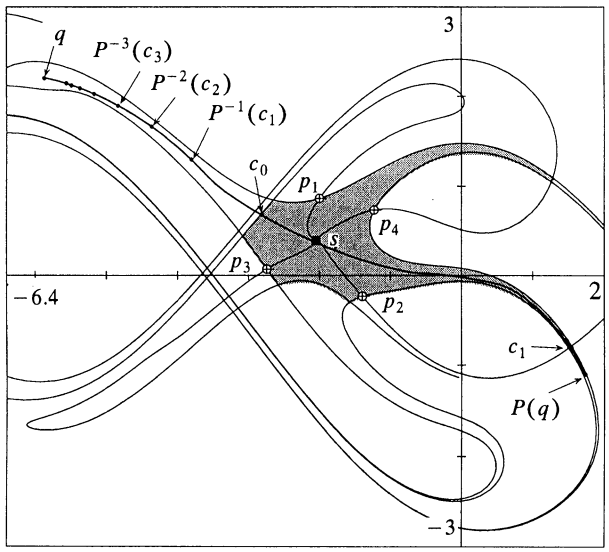


Figure 8. A basin cell; the points $P^{-i}(c_i)$ illustrates the proof of Theorem 2.

the stable separatrix of the saddle. Thus the stable separatrix is part of the accessible boundary.

Theorem 2. *The accessible boundary of B_k is exactly the union of the stable separatrices $\sigma^+(p_{k,i})$ $i = 1, \ldots, 4$.*

The proof consists of looking at Figure 8.

The colored neighborhood C_k of the sink s_k (called a *basin cell* in [**12**]) is bounded by arcs of four stable separatrices $\sigma^+(p_{k,i})$ and arcs of the four unstable separatrices $\sigma^-(p_{k,i})$, which except for endpoints are contained in the interior of the basin. Thus any accessible boundary point q of B_k not in $\bigcup_i \sigma^+(p_{k,i})$ is necessarily outside C_k, and a path

$$\gamma : [0,1] \mapsto \overline{B}_k, \gamma([0,1]) \subset B_k$$

joining q to s_k intersects one of these four arcs, in points c_0. Similarly, the path $P^m(\gamma)$ intersects one of these arcs in a point c_m. The points $z_m = P^{-m}(c_m)$ must be on γ, and must converge to q since for any $\epsilon > 0$, the set $\gamma([0, 1 - \epsilon])$ is a compact subset of B_k. Thus $P^m(\gamma([0, 1 - \epsilon]))$ is inside C_k (or any neighborhood of s_k for m sufficiently large).

But the c_m lie in four compact arcs of $\bigcup_i \sigma^-(p_{k,i})$, hence $P^{-m}(c_m)$ is very close to one of the saddles for m large. So q is one of the saddles $p_{k,i}$, and hence is on its stable separatrix.

This ends the proof of Theorem 2 (or at least a fairly convincing argument; it is not a rigorous proof, as we discuss in Sections 11 and 13); now to justify Theorem 1.

First, it is enough to show that each accessible point of ∂B_0 (the boundary of B_0) can be approached by every other basin. Indeed, every point of ∂B_0 can be approached by accessible points, so if we can show that each accessible point of ∂B_0 is in the boundary of every other basin, then every point of ∂B_0 is in the boundary of every other basin.

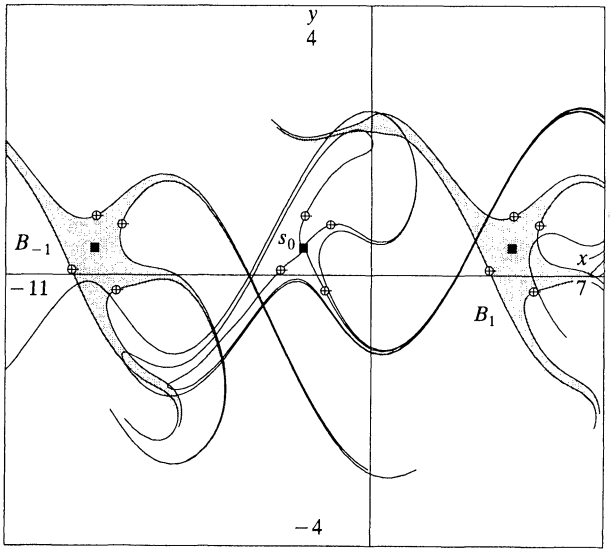


Figure 9. All four of the unstable separatrices from the points $p_{0,i}$ enter both B_1 and B_{-1}.

Second, it is enough to know that the four *outward pointing branches* of the unstable separatrices for the four accessible saddles in ∂B_0 enter every basin. Indeed, if the four unstable separatrices $\sigma^-(p_{0,i})$, for $i = 1, 2, 3, 4$, enter B_n, then the inverse images of B_n accumulate to $p_{0,i}$, hence to the entire stable separatrix $\sigma^+(p_{0,i})$. This shows a little more: if all four $\sigma^-(p_{0,i})$ enter B_n, then no curve can enter B_0 without crossing a stream of B_n, i.e., entering B_n.

Third, rather than show that the outward-pointing part of each $\sigma^-(p_{0,i})$ enters all the basins B_n, for n any integer, it is enough to show that it enters the two neighboring basins B_1 and B_{-1}. We can prove this by induction. Figure 9 shows that the four separatrices $\sigma^-(p_{0,i})$, $i = 1, 2, 3, 4$ enter the basins B_{-1} and B_1.

Now suppose they enter B_k for some $k > 1$. But they cannot enter B_k without entering B_{k+1}, because the $\sigma^-_{k,i}$ enter B_{k+1}, so that their inverse images give streams of B_{k+1}, which they must ford to enter B_k.

8. SOLUTIONS NOT ATTRACTED TO THE SINKS. In this section we use techniques mainly due to Smale [**14**] to show that the differential equation for our pendulum has trajectories that carry out any specified sequence of gyrations. During one time interval $I_k = [2k\pi, 2(k+1)\pi)$ a solution $(x(t), x'(t))$ may satisfy $x(t) = 0 \pmod{2\pi}$ exactly

$[-1]$ once with $x' < 0$,

$[0]$ never,

$[1]$ once with $x' > 0$,

$[\mathrm{NA}]$ none of the above.

These events correspond to the pendulum crossing the downward position exactly once clockwise, not crossing it, crossing it once counterclockwise, or doing something else. In particular, the attracting solutions belong to the "none of the above" category, because they cross the downward position twice during each period. So, eventually, do all solutions that are attracted to them. Thus Theorem 3 describes solutions entirely contained in the separator, which are never attracted to one of the sinks.

Theorem 3. *Given any bi-infinite sequence of events* $\ldots E_{-1}, E_0, E_1, \ldots$ *with* $E_k \in \{[-1], [0], [1]\}$ (*but not* $[NA]$), *there exists a solution of our differential equation that during each time interval* $[2k\pi, 2(k+1)\pi)$ *will "do"* E_k.

Thus given any sequence of gyrations one might choose, there is a solution that does exactly that. In particular, any sequence (E_i) of period m and that sum to 0 over one such period corresponds to a periodic cycle of period m for P. Theorem 3 is very similar to Alekseev's theorem, and is proved the same way: by exhibiting a Smale horseshoe. In Alekseev's case this requires a delicate perturbation argument; we show how the computer can make such a result transparent.

We have found a sequence of fixed sinks s_k that correspond to the downward equilibrium of the unforced pendulum. There is also a sequence of fixed saddles corresponding to a periodic solution of the original differential equation of period 2π near the unstable upward equilibrium. If you draw a sequence of quadrilaterals Q_k roughly aligned with the stable and unstable separatrices of these fixed saddles, as in Figure 10, you expect the image of such a quadrilateral to be compressed in the stable direction and stretched in the unstable direction, becoming long and filiform.

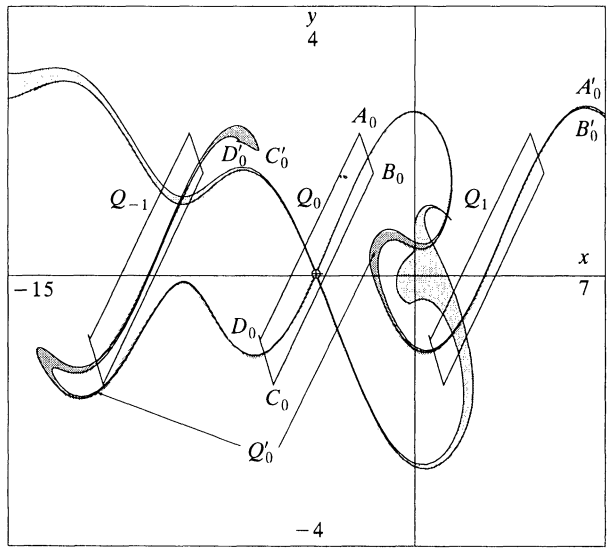


Figure 10. The quadrilaterals Q_{-1}, Q_0, Q_1, together with the forward and backwards images of Q_0.

We now describe the set of points

$$Q_k(E_0, E_1, \ldots, E_N) = \{p | P^n(p) \in Q_{k+E_0+\cdots+E_{n-1}} \text{ for } 0 \le n \le N\}.$$

Let A_0, B_0, C_0, D_0 denote the corners of Q_0, as shown in Figure 10. The set $P(Q_0)$ is the curvilinear quadrilateral Q_0', shaded in Figure 10, with vertices A_0', B_0', C_0', D_0'. The key property of the image is that it crosses the quadrilaterals Q_1 and Q_{-1}, as well as itself, in each case going from top to bottom (or bottom to top), with the top A_0B_0 and bottom C_0D_0 mapping outside these quadrilaterals.

This implies that each of $Q_0([-1]), Q_0([0]), Q_0([1])$ forms a full-width subrectangle of Q_0. Figure 11 shows the forward and backwards images of Q_0, Q_{-1} and Q_1, and a blow-up of showing how these intersect Q_0. Indeed the backwards images (light shading) form full-width subrectangles. Of course, Q_1 and Q_{-1} also contain such subrectangles $Q_1(E_0)$, etc. The inverse image $P^{-1}(Q_{E_0}(E_1))$ is then again a (thinner) full-width subrectangle $Q_0(E_0, E_1)$.

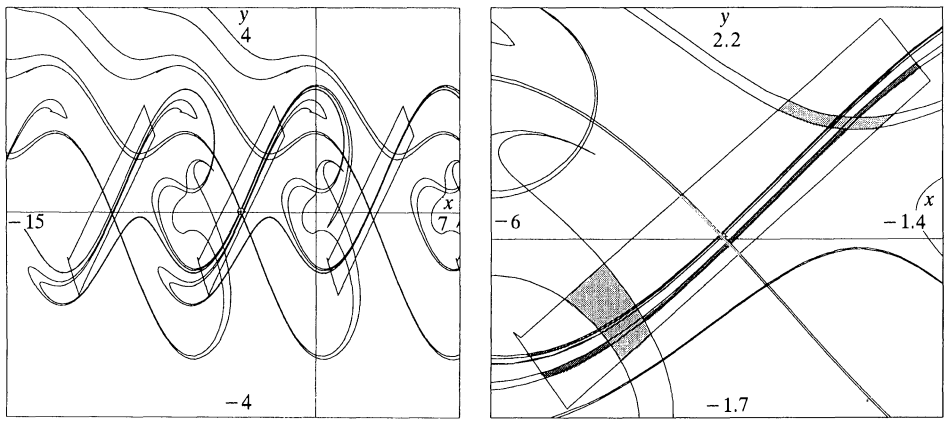


Figure 11. The forward images of Q_{-1}, Q_0, Q_1, and their intersections with Q_0. At right a blow-up of Q_0.

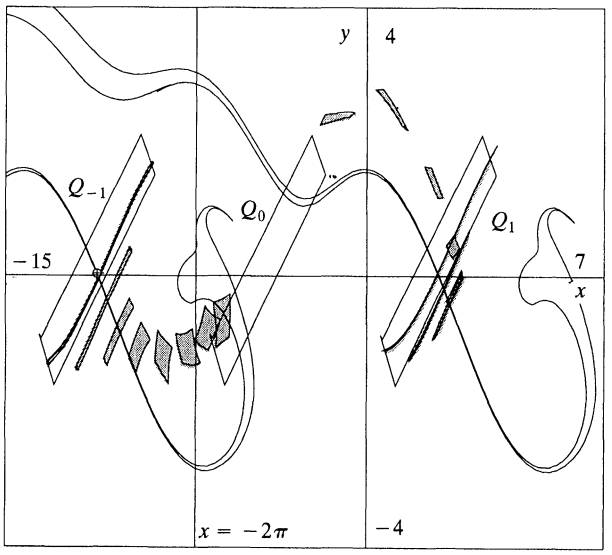


Figure 12. How the quadrilaterals move during one period.

Continuing this way, we see that for any finite sequence $(E_0, E_1, \ldots, E_N)$, the corresponding set $Q_0(E_0, E_1, \ldots, E_N)$ is a full-width subrectangle of Q_0. Finally, the assignment of an infinite forward trajectory restricts the initial position to an infinite intersection of nested full-width subrectangles of Q_0; such an intersection is a connected subset of Q_0 connecting one side of Q_0 to the other. In fact, it is a smooth curve, but this requires writing some inequalities.

A similar argument shows that any finite backwards trajectory restricts the final position to a full-height subrectangle of Q_0, and an infinite backwards trajectory leads to a connected subset joining A_0B_0 to C_0D_0 (again in fact a smooth curve). If $X, Y \subset Q_0$ are connected subsets, with X joining D_0A_0 to B_0C_0 and Y joining A_0B_0 to C_0D_0, then $X \cap Y \neq 0$. Thus there is a point realizing any prescribed symbolic trajectory.

Finally, I claim that the points of $Q_0([-1]), Q_0([0]), Q_0([1])$ realize the events $[-1], [0]$, and $[1]$, respectively. Figure 12 shows the images of $Q_0(+1)$ and $Q_0(-1)$ at times

$$0, \frac{2\pi}{8}, \frac{4\pi}{8}, \frac{6\pi}{8}, \frac{8\pi}{8}, \frac{10\pi}{8}, \frac{12\pi}{8}, \frac{14\pi}{8}, \frac{16\pi}{8}.$$

The first set certainly seems to cut the line $x = 2\pi$ exactly once with $y > 0$; the second set seems to cut the line $x = 0$ once with $y < 0$.

9. CONTROLLING THE PENDULUM. Imagine that the pendulum is massive, and is being used as a flywheel to control some very delicate operation, like polishing the mirror of a telescope. An array of lasers is constantly monitoring the operation, deciding on the fly whether the pendulum should turn clockwise, counterclockwise, or wait until the mirror has been repositioned.

The previous section showed that there are motions of the pendulum performing any specified sequence of gyrations, in particular the one required a posteriori by the polisher. But on second thought this seems useless: these motions are extremely unstable, and the slightest error in the initial condition destroys them, as

well as any perturbation of the differential equation itself. But if the machine is to perform any work, this inevitably perturbs the differential equation, in a way that is essentially unpredictable (you cannot predict how much work one swipe of the polisher will accomplish), and in any case we don't know ahead of time the sequence of swipes and stops the task will require.

On third thought, we see that the instability of the specified motions is exactly what should make them useful! Suppose that our array of sensors controls the current $f(t)$ that is forcing the pendulum, changing it from $\cos(t)$ to something like

$$(1 + a(t))\cos(t) \quad \text{(amplitude modulation)} \quad \text{or}$$

$$\cos((1 + a(t))t) \quad \text{(frequency modulation)},$$

where $a(t)$ represents the fine-tuning necessary to achieve the desired sequence of gyrations. The point is that we do not have to figure out what sequence we want ahead of time: the sensors can react to the polishing of the telescope on the fly, computing the adjustment $a(t)$ that is necessary. It is because of the instability that you can keep $a(t)$ small and still realize any sequence of gyrations: you don't need to grind to a halt, compute, and start up again; the corrections can be done smoothly. A useful analogy is skiing: a beginning skier plants his skis well apart, seeking stability, which is fine until he tries to turn and discovers he can't. An expert skier, with skis parallel and touching, is highly unstable, and a slight wiggle of the hips allows him to negotiate a mogul. Of course he doesn't plot his entire path at the top of the mountain; he calculates the slight adjustments $a(t)$ as they are needed.

Theorem 4. *For any sequence of events $E_0, E_1, \ldots$ and any sufficiently small disturbance $b(t)$ of the forcing term $\cos t$, there exists a function $a(t)$ of the same order of magnitude as $b(t)$ and an initial condition $x(0), x'(0)$ such that the solution of the differential equation*

$$x'' + 0.1x'(t) + \sin(x) + b(t) = (1 + a(t))\cos t$$

with those initial conditions realizes the specified sequence of events.

This result is fairly obvious: choose $a(t)$ as the pendulum approaches the upwards position so as to speed it up or slow it down as required. The problem is how to compute the $a(t)$, in terms of available data. Clearly $a(t)$ should depend only on the values of b up to time $t - 2\pi$; it should not depend on the specified sequence of events very far ahead, as this is unknown. How small can $a(t)$ be made? How far ahead in the required sequence of events does it need to look? How sensitive is it to small errors in the sensors? ...

10. CONTROL AND CELESTIAL MECHANICS. To return to celestial mechanics for a moment, it is interesting to note that when sending a spaceship to visit the outer solar system, NASA uses the instabilities of the differential equations describing gravity in much the same way as we have used the instabilities of the pendulum. It is well beyond present-day engineering to send a spaceship out of the solar system by simply using its fuel to accelerate it. Instead, it is allowed to "fall" into the sun, with an orbit that passes close to Venus. It then loops around Venus; we can imagine that it is the "satellite" in the three-body system consisting of itself, Venus, and the sun.

This system is similar to Alekseev's (somewhat more complicated: a Poincaré section would need to be 4-dimensional rather than 2), and one can prescribe an orbit so that the space ship steals a tiny amount of potential energy from Venus, speeding up enormously in the process, and ends up in a very unstable state where a small push by guidance rockets can put it on the path to Jupiter.

This scenario is then repeated near Jupiter, Saturn, and Uranus, with the spaceship each time gaining momentum, and using small pushes to head itself in the direction of the next destination. Thus the chaos of the solar system is essential to its exploration.

11. WHAT IS PROVED? To what extent does this paper prove anything? As written, no statement is proved anywhere: for the punchline we just looked at a computer picture. How do we know that these pictures are right? I do not address the possibility that the programs have essential bugs and are computing something other than what I think, or the esoteric possibility that the computer arithmetic is wrong. But even if the computer is computing exactly what I think, that is still only an approximation to solutions of the differential equations; we need to quantify the quality of the approximation. The contribution of round-off error also should be addressed.

Actually, many of the results are not hard to prove rigorously, namely all those where we have to show that after time 2π, solutions are within some fairly large ϵ of the value suggested by the computer drawings.

Good estimates of long-term errors of numerical approximations to solutions of differential equations are notoriously hard to come by, but that is not really a problem here. First, we do not need *good* estimates (solutions need only be accurate to about 0.1); second, the time considered is *not long* (2π); and most important, the differential equation has a small *Lipschitz constant* ($\sqrt{2.001} < 1.42$). Errors in solutions to differential equations grow at most exponentially, at a rate e^{kt}, where t is time (in our case, 2π) and k is the Lipschitz constant; with $k < 1.42$, errors grow at a fairly small interest rate, and can be controlled for a short time.

Using these numbers, a straightforward computation using the *fundamental inequality* ([**8**, Chapters 4 and 6]) shows that if the initial velocity satisfies $|x'(0)| < 3$, then Euler's method with step-length $h = 0.000002$ gives results accurate to 0.1 after time 2π. Moreover, the same inequality shows that round-off error contributes a much smaller error yet. This is not a good way to do such numerics; better numerical methods give much better estimates [**5**]. For instance, formula (14) of [**2**] can be used to show that the fourth order Runge-Kutta method with step 0.005 has more than the needed precision.

A word of caution, though. The elementary bound above says that errors of all types are multiplied by at most $e^{2\pi 1.42} \approx 7\,500$ over one time period. It is not too difficult to improve this to $e^{2\pi 1.1} \approx 1\,000$, and one could improve it further. But one could not improve it very much further.

Consider for example the completely unavoidable error caused by the computer's inability to handle numbers with infinite precision. If it handles numbers to 16 significant digits, you may think you are starting at a saddle point, but your initial error (the distance between the saddle point and where you really are) may be as great as 10^{-16}. The largest eigenvalue λ of the linearization of P at the fixed saddles in the Q_k is about 321 (according to the computer). As long as you are in the region where P is approximately its linearization at this saddle, errors of all types are expanded by a factor of λ over one time period, and hence λ^m over m

time periods. So after m iterations the error will have mushroomed to $10^{-16}(321^m)$: for $m = 7$ the initial minute error will have grown to 35. But already for an error of 1, you will have been booted out of the region where the linearization is a reasonable approximation to reality.

Thus no numerical method can guarantee even one digit of accuracy after six time periods, if we are computing with 16 significant digits. In fact, the reality is much worse than that, and I wouldn't trust anything after four time periods without some good reason.

12. A POSTERIORI BOUNDS. Good reasons to trust solutions are available: I advocate extrapolation, as described in [**8**, Chapter 3]. At the moment, this works only for fixed step-length, but for a Poincaré mapping of a differential equation, fixed step-length is probably best anyway. For other possible methods, consult [**10**].

Denote by $u_h(t)$ the numerical approximation to the solution of some differential equation given by the standard fourth order Runge-Kutta method, with $u_h(0) = a$. Then the theory asserts that for each fixed t the approximation $u_h(t)$ converges to the value of the solution $u(t)$, and that we have an asymptotic development

$$u_h(t) = u(t) + Ch^4 + o(h^4).$$

The exponent 4 is a feature of this approximation procedure; other procedures have different exponents.

If for some h we know $u_h(t), u_{h/2}(t)$, and $u_{h/4}(t)$, and we assume that we have an asymptotic development of the form $u_h(t) = u(t) + Ch^k + o(h^k)$ for some k, we can extrapolate the values of k and of C from the values of the approximate solutions:

$$k = \frac{1}{\log 2}\log\left|\frac{u_h(t) - u_{h/2}(t)}{u_{h/2}(t) - u_{h/4}(t)}\right| \quad \text{and} \quad C = \frac{2^k}{2^k - 1}\frac{u_h(t) - u_{h/2}(t)}{h^k}.$$

Now suppose we calculate $u_{h/2^m}(t)$ for a range of values of m, focusing on the expression for k above. The theory says that as m increases, the value of k should approach 4, but that doesn't take round-off error into account; typically the value of k approaches 4 as m increases, then veers away from 4 as round-off error takes over. If there is a range of values of m where k is close to 4, the approximation is happening the way the theory predicts, and we can probably trust the corresponding estimate of the error. The following data illustrates this for our differential equation, solved for $0 \le t \le 16\pi$, i.e., for 8 periods. We start with the two initial positions (7.15859, 0.14097) and (7.16859, 0.14097). The extrapolations we find are

	first solution		second solution	
steps	order	error	order	error
6				
12		22.45		86.22
24	3.07	2.67	1.05	41.48
48	−1.79	9.31	2.61	6.77
96	3.26	0.96	−0.15	6.84
192	−0.44	1.31	−1.09	14.64
384	−2.01	5.27	3.02	1.80
768	−0.06	5.48	4.96	0.057
1536	5.13	0.16	4.19	0.003

Thus, the first approximation never becomes reliable; the order is never close to 4. In particular, there is no reason to think that the quantity in the "error" column is actually an estimate of the error. But the second appears to be converging nicely, with the order approaching 4, and probably the error estimate of 0.003 is reliable. Thus although any estimate we make *a priori* for a bound for the error is bound to be wildly pessimistic, after the computation we can make a good guess as to how reliable it is.

13. QUESTIONS AND OBSERVATIONS.

(1) Are there any periodic sinks other than the attracting fixed points we found? I have no idea how to attack this problem. For one thing, I don't trust computer drawings on this point: in many instances I eventually found sinks whose basins were too small to be visible on computer drawings unless you knew where to look. For another, the answer might depend in the most delicate way on the parameters: there definitely are other attracting fixed points when the forcing term is $1.22 \cos t$ instead of $\cos t$; for example, there is a sink of period 3, where solutions go from the point with coordinates $x = -1.29785$, $y = 1.0025$ to the point $x = -1.3349$, $y = -0.21286$, to the point $x = -3.004469$, $y = 0.17586$, and then back to the first point.... In fact, with those parameters there are at least two more sinks of period 3, in addition to all the translates of the three sinks by 2π.

This problem may be unsolvable. John Milnor's candidate for the simplest unsolvable problem of mathematics is the question: "Does the polynomial $x^2 - 1.5$ have an attracting cycle?" Of course, if it does, one can find it with a finite amount of work. But if it doesn't, there may be no proof of this fact.

(2) Is the complement of all the basins B_k of measure 0? This would mean that with probability 1 every initial point is attracted to a sink. I think this is the case, but have no solid grounds for this belief. Even the computer isn't very definite, and besides, this is one point where numerical error might really be important: the perturbations of the period mapping due to errors of integration and round-off might affect the probability of being attracted to a sink.

ACKNOWLEDGMENTS. I thank Stan Wagon, Mukund Thattai, and Robert Terrell for their many helpful comments and suggestions.

REFERENCES

1. V. M. Alekseev, On the capture orbits for the three-body problem for negative energy constant, *Uspekhi Mat. Nauk* 24 (1969) 185–186.
2. L. Bieberbach, On the remainder of the Runge-Kutta formula in the theory of ordinary differential equations, *Zeit. angewandte Math. Physik* 2 (1951) 233–248.
3. R. Borelli and C. Coleman, Computers, lies and the fishing season, *College Math. J.* 25 (1994) 401–412.
4. L. E. J Brouwer, Zur Analysis Situs, *Math. Annalen* 68 (1910) 422–434.
5. P. Henrici, *Discrete variable methods in ordinary differential equations*, Wiley, New York, 1962.
6. B. Hinkle, J. Hubbard, and B. West, *Planar Systems* and *Planar Iterations*, to be published by Springer Verlag. These programs constitute the latest version of *MacMath* by Hubbard and West, originally published by Springer Verlag in 1991.

7. J. Hubbard and R. Oberste-Vorth, Hénon mapping in the complex domain II: Projective and inductive limits of polynomials, *Proceedings of the NATO summer conference in Hillerod*, 1993.
8. J. Hubbard and B. West, *Differential Equations, a dynamical systems approach*, Vol. I, Springer Verlag, New York, 1992 and Vol. II, Springer Verlag, New York, 1994.
9. J. Kennedy and J. Yorke, Basins of Wada, *Physica D* 51 (1991) 213–255.
10. R. Knapp and S. Wagon, Check your answers... but how?, *Mathematica in Education and Research*, 7–4.
11. J. Moser, Stable and random motions in dynamical systems, *Annals of Mathematics Studies*, Princeton University Press, Princeton, N.J., 1973
12. H. Nusse and J. Yorke, Wada basin boundaries, March 1994, preprint.
13. D. Schwalbe and S. Wagon, *VisualDSolve: Visualizing Differential Equations with Mathematica*, Springer/TELOS, N.Y. 1997.
14. D. Smale, Diffeomorphisms with many periodic points, *Differential and combinatorial topology*, edited by S. S. Cairns, Princeton University Press, 1965, pp. 63–80.
15. K. Yoneyama, Theory of continuous set of points, *Tohoku Math. J* 11–12 (1917) 43.

JOHN H. HUBBARD, professor of mathematics at Cornell, received his BA from Harvard and his Ph.D. from the University of Paris. He has been a strong believer in using computers in mathematics since he made the first pictures of the Mandelbrot set on his first computer, an Apple 2. He believes that teaching and research enrich each other. His research interests include complex analysis, Teichmüller theory, and differential equations.
Cornell University, White Hall, Ithaca, NY 14853
jhh8@cornell.edu

NOTES

Edited by **Jimmie D. Lawson and William Adkins**

The Hyperbolic Pythagorean Theorem in the Poincaré Disc Model of Hyperbolic Geometry

Abraham A. Ungar

Sometime in the sixth century B.C. Pythagoras of Samos discovered the theorem that now bears his name in Euclidean geometry. The extension of the Euclidean Pythagorean theorem to hyperbolic geometry, which is commonly known as the hyperbolic Pythagorean theorem (see [**3, 5, 6, 9–11**]), does not have a form analogous to the Euclidean Pythagorean theorem, so some authors have concluded that a truly hyperbolic Pythagorean theorem does not exist. For example, Wallace and West assert "the Pythagorean theorem is strictly Euclidean" since "in the hyperbolic [Poincaré disc] model the Pythagorean theorem is not valid!" [**15**]. We show that a natural formulation of the hyperbolic Pythagorean theorem does exist: it expresses the square of the hyperbolic length of the hypotenuse of a hyperbolic right angled triangle as a natural "sum" of the squares of the hyperbolic lengths of the other two sides.

The most general Möbius transformation of the complex unit disc $D = \{z : |z| < 1\}$ in the complex z-plane [**2, 4, 8**],

$$z \mapsto e^{i\theta}\frac{z_0 + z}{1 + \bar{z}_0 z} = e^{i\theta}(z_0 \oplus z), \tag{1}$$

defines the *Möbius addition* $\oplus$ in the disc, which allows the Möbius transformation of the disc to be viewed as a *Möbius left translation*

$$z \mapsto z_0 \oplus z = \frac{z_0 + z}{1 + \bar{z}_0 z}$$

followed by a rotation. Here $\theta \in \mathbb{R}$ is a real number, $z_0 \in D$, and $\bar{z}_0$ is the complex conjugate of z_0. A left Möbius translation is also called a left *gyrotranslation* [**13**]. Left gyrotranslations occur frequently in hyperbolic geometry [**7**, p. 55]. and are sometimes called *hyperbolic pure translations* [**9**, p. 224].

The prefix *gyro* that we use to emphasize analogies stems from the *Thomas gyration*, which results, in turn, from the abstraction of the relativistic effect known as the *Thomas precession* [**13, 14**]. The relevance of the Thomas precession to hyperbolic geometry is not unexpected [**9**, p. 251] since this geometry underlies relativistic velocities. The sensitivity of Thomas precession to the non-Euclidean nature of the geometry of spacetime has attracted NASA's interest in measuring the Thomas precession of gyroscopes of unprecedented accuracy in Earth orbit; see http://einstein.stanford.edu.

The *Poincaré hyperbolic distance function* in D is [**2**]

$$d(a,b) = \left| \frac{a-b}{1-\bar{a}b} \right| = |a \ominus b|, \tag{2}$$

where we use the obvious notation $a \ominus b = a \oplus (-b)$ for $a, b \in D$. It satisfies the Möbius triangle inequality

$$d(a,c) \leq d(a,b) \oplus d(b,c), \tag{3}$$

which involves the Möbius addition $\oplus$ of two real numbers in the complex unit disc D. We prove (3) after the proof of our main theorem and a discussion of some relevant group theoretic properties of Möbius addition. The right hand side of (3) can be written as

$$\tanh\left(\tanh^{-1} d(a,b) + \tanh^{-1} d(b,c)\right) \tag{4}$$

so that the Möbius triangle inequality can be written as an inequality

$$\tanh^{-1} d(a,c) \leq \tanh^{-1} d(a,b) + \tanh^{-1} d(b,c) \tag{5}$$

that involves the ordinary, rather than the Möbius, addition of real numbers. The hyperbolic distance function in D is commonly defined in the literature by [**7**, p. 53]

$$h(a,b) = \tanh^{-1} d(a,b) = \frac{1}{2} \ln \frac{1+d(a,b)}{1-d(a,b)} \tag{6}$$

rather than by $d(a,b)$ in which case we have in the triangle inequality

$$h(a,c) \leq h(a,b) + h(b,c) \tag{7}$$

for all $a, b, c \in D$. The complex unit disc with its Poincaré distance function, called the *Poincaré disc*, gives the Poincaré disc model of hyperbolic geometry, in which geodesic lines are circular arcs that intersect the boundary of the disc orthogonally [**3**].

Theorem. (The Hyperbolic Pythagorean Theorem) *Let* $\triangle abc$ *be a hyperbolic triangle in the Poincaré disc, whose vertices are the points* a, b *and* c *of the disc and whose sides (directed counterclockwise) are* $A = -b \oplus c$, $B = -c \oplus a$, *and* $C = -a \oplus b$. *If the two sides* A *and* B *are orthogonal, then* $|A|^2 \oplus |B|^2 = |C|^2$.

Proof: Let $\triangle abc$ be any hyperbolic triangle whose vertices are the points a, b, and c of the disc, and whose sides, A, B, and C, are geodesic segments that join the vertices, as shown in Figure 1. The measure of the hyperbolic angle between two sides of a hyperbolic triangle is given by the Euclidean measure of the angle formed by Euclidean tangent rays [**3**]. A *hyperbolic right triangle* is a hyperbolic triangle one of whose angles is $\pi/2$. Furthermore, let $\triangle abc$ be a hyperbolic right triangle whose sides A and B are orthogonal. Its right angle can be moved to the center of D by an appropriate Möbius transformation (1) such that its two orthogonal sides lie on the real and on the imaginary axes of D, as shown in Figure 1. Möbius transformations of the disc preserve both the hyperbolic length of geodesic segments and the measure of hyperbolic angles. Hence, the resulting triangle $\triangle a'b'c'$, obtained by moving $\triangle abc$ as shown in Figure 1, is congruent to $\triangle abc$ in the sense that the two triangles $\triangle a'b'c'$ and $\triangle abc$ possess equal hyperbolic lengths for corresponding sides and equal measures for corresponding angles.

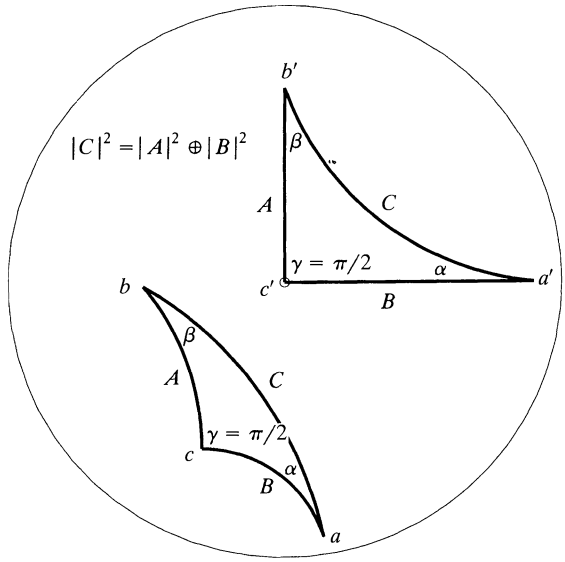


Figure 1. *The Hyperbolic Pythagorean Theorem in the complex unit disc.* The square of the hyperbolic length of the hypotenuse of a hyperbolic right triangle equals the Möbius sum of the squares of the hyperbolic lengths of the other two sides. Furthermore, $\sin \alpha = \gamma |A|/(\gamma_C |C|)$ and $\sin \beta = \gamma_B |B|/(\gamma_C |C|)$.

The vertices of the relocated hyperbolic right triangle $\triangle a'b'c'$ are $a' = x$, $b' = iy$, and $c' = 0$, for some $x, y \in (-1, 1)$. The hyperbolic length of the geodesic segment joining two points a and b of the disc is $d(a, b) = |b \ominus a|$. Accordingly, the hyperbolic lengths of the sides A, B, C of the triangle $\triangle a'b'c'$ are $|A|$, $|B|$, and $|C|$ given by

$$\begin{aligned} |A|^2 &= |b' \ominus c'|^2 = y^2, \\ |B|^2 &= |a' \ominus c'|^2 = x^2, \quad \text{and} \\ |C|^2 &= |a' \ominus b'|^2 = |x \ominus iy|^2 = \left| \frac{x - iy}{1 - ixy} \right|^2 = x^2 \oplus y^2. \end{aligned} \tag{8}$$

Hence

$$|A|^2 \oplus |B|^2 = |C|^2, \tag{9}$$

which verifies the hyperbolic Pythagorean theorem for hyperbolic right triangles in the Poincaré disc. ■

The Hyperbolic Pythagorean Theorem is not an isolated analogy with Euclidean geometry; analogies between the Poincaré disc model of hyperbolic geometry and Euclidean plane geometry abound in gyrogroup theory {**12**]. It is shown there that the Möbius addition, $\oplus$, is analogous to the common vector addition, $+$, in Euclidean plane geometry. If we define

$$\operatorname{gyr}[a; b] = \frac{a \oplus b}{b \oplus a} = \frac{1 + a\bar{b}}{1 + \bar{a}b}, \tag{10}$$

then $\operatorname{gyr}[a; b]$ has modulus 1 and for all $a, b, c \in D$ the following group-like properties of $\oplus$ can be verified by straightforward algebra:

$a \oplus b = \operatorname{gyr}[a; b](b \oplus a)$	Gyrocommutative Law
$a \oplus (b \oplus c) = (a \oplus b) \oplus \operatorname{gyr}[a; b]c$	Left gyroassociative Law
$(a \oplus b) \oplus c = a \oplus (b \oplus \operatorname{gyr}[b; a]c)$	Right gyroassociative Law
$\operatorname{gyr}[a; b] = \operatorname{gyr}[a \oplus b; b]$	Left Loop Property
$\operatorname{gyr}[a; b] = \operatorname{gyr}[a; b \oplus a]$	Right Loop Property

A resulting geometrically important identity, also verifiable by straightforward algebra, is [**12**]

$$(x \oplus a) \ominus (x \oplus b) = \operatorname{gyr}[x, a](a \ominus b) \tag{11}$$

for all $a, b, x \in D$. Taking the modulus of each side of (11) gives

$$d(x \oplus a, x \oplus b) = d(a, b), \tag{12}$$

which shows that the Poincaré distance function (2) is invariant under Möbius left gyrotranslations.

To verify the Möbius triangle inequality (3), let $\gamma_a = (1 - |a|^2)^{-1/2}$ for any $a \in D$. Then $\gamma_a = \gamma_{|a|}$ is a monotonically increasing function of $|a|$ that satisfies the useful identity

$$\gamma_{a \oplus b} = \gamma_a \gamma_b |1 + \bar{a}b| \tag{13}$$

for all $a, b \in D$ [**1**, p. 2], as one can verify by squaring both sides.

It follows from (13) that

$$\gamma_{||a| \oplus |b||} = \gamma_{|a| \oplus |b|} = \gamma_{|a|}\gamma_{|b|}(1 + |a|\,|b|) \geq \gamma_a \gamma_b |1 + \bar{a}b| = \gamma_{a \oplus b} = \gamma_{|a \oplus b|}. \tag{14}$$

Since $||a| \oplus |b|| = |a| \oplus |b|$, and since $\gamma_z = \gamma_{|z|}$ is a monotonically increasing function of $|z|$, the inequality in (14) implies the inequality

$$|a| \oplus |b| \geq |a \oplus b| \tag{15}$$

for all $a, b \in D$.

Replacing x by $-x$ in (11), and noting that $-(-x \oplus b) = x \ominus b$, we have

$$(-x \oplus a) \oplus (x \ominus b) = \operatorname{gyr}[-x, a](a \ominus b) \tag{16}$$

for all $x, a, b \in D$. Finally, (16) and (15) imply

$$\begin{aligned} d(a, b) &= |a \ominus b| = |\operatorname{gyr}[-x, a](a \ominus b)| = |(-x \oplus a) \oplus (x \ominus b)| \\ &\leq |-x \oplus a| \oplus |x \ominus b| = d(a, x) \oplus d(x, b) \end{aligned}$$

for all $a, b, x \in D$, which proves the Möbius triangle inequality (3).

REFERENCES

1. Lars V. Ahlfors, *Conformal Invariants Topics in Geometric Function Theory*, McGraw-Hill. New York, 1973.
2. Stephen D. Fisher, The Möbius group and invariant spaces of analytic functions, *Amer. Math. Month.* **95** (1988) 514–527.
3. *Marvin J. Greenberg, Euclidean and non-Euclidean geometries: development and history*, 2nd ed., Freeman, San Franscisco, 1980.
4. Robert E. Green and Steven G. Krantz, *Function Theory of One Complex Variable*, Pure Appl. Math, A Wiley-Interscience Ser. Wiley, New York, 1997, p. 185.
5. Michael Henle, *Modern Geometries: The Analytic Approach*, Prentice Hall, NJ, 1997, p. 102.
6. David C. Kay, *College Geometry*, Holt, Reinhart and Winston, New York, 1969, p. 317.

7. Steven G. Krantz, *Complex Analysis: The Geometry Viewpoint*, Carus Mathematical Monographs, 23, Math. Assoc. of Amer., Washington, DC, 1990.
8. Serge Lang, *Complex Analysis*, 3rd ed. Springer, New York, 1993, p. 213.
9. Arlan Ramsay and Robert D. Richtmyer, *Introduction to Hyperbolic Geometry*, Springer, New York, 1995.
10. Hans Schwerdtfeger, *Geometry of Complex Numbers, Circle Geometry, Möbius Transformations, Non-Euclidean Geometry*, Dover, New York, 1979, p. 146.
11. William P. Thurston, *Three-Dimensional Geometry and Topology*, Vol. 1, ed. by Silvio Levy, Princeton Univ. Press, New Jersey, 1997, p. 81.
12. Abraham A. Ungar, The holomorphic automorphism group of the complex disk, *Aequat. Math.* **47** (1994) 240–254.
13. Abraham A. Ungar, Thomas precession: its underlying gyrogroup axioms and their use in hyperbolic geometry and relativistic physics, *Found. Phys.* **27** (1997) 881–951.
14. Abraham A. Ungar, From Pythagoreas to Einstein: The Hyperbolic Pythagorean Theorem, *Found. Phys.* **28** (1998) 1283–1321.
15. Edward C. Wallace and Steven F. West, *Roads to Geometry*, 2nd ed., Prentice Hall, NJ, 1998, pp. 362–363.

North Dakota State University, Fargo, North Dakota 58105
ungar@plains.NoDak.edu

Is the Composite Function Integrable?

Jitan Lu

It is well known that the composition of two continuous functions is continuous and hence Riemann integrable. However, the composition of two Riemann integrable functions may or may not be Riemann integrable. For example, let

$$f(y) = \begin{cases} 1 & \text{when } y \neq 0, \\ 0 & \text{when } y = 0, \end{cases}$$

and

$$g(x) = \begin{cases} 0 & \text{when } x \text{ is an irrational number,} \\ \frac{1}{p} & \text{when } x = \frac{q}{p}, \text{ where } p \text{ and } q \text{ are two coprime integers.} \end{cases}$$

Then

$$f \circ g(x) = \begin{cases} 0 & \text{when } x \text{ is an irrational number,} \\ 1 & \text{when } x = \frac{q}{p}, \text{ where } p \text{ and } q \text{ are two coprime integers.} \end{cases}$$

Both f and g are Riemann integrable on $[0, 1]$, but the composition $f \circ g$ is not. Therefore, it is natural to ask whether the composition of two functions is still Riemann integrable, when one is Riemann integrable and the other is continuous.

In what follows, we let f be a function defined on the interval $[a, b]$, and let g be a function defined on the interval $[c, d]$ with its range contained in $[a, b]$.

Question 1. If f is continuous on $[a, b]$ and g is Riemann integrable on $[c, d]$, is the composition $f \circ g$ Riemann integrable on $[c, d]$?

The answer is yes.

Since f is continuous on the closed interval $[a, b]$, it is uniformly continuous on $[a, b]$. Hence, for each $\varepsilon > 0$, there exists a $\delta > 0$, such that for any ξ_1 and ξ_2 in $[a, b]$ with $|\xi_1 - \xi_2| < \delta$ we have

$$|f(\xi_1) - f(\xi_2)| < \frac{\varepsilon}{2(d-c)}. \tag{1}$$

Moreover, f is bounded on $[a, b]$; say, $|f(y)| \leq M$ for all $y \in [a, b]$.

Since g is Riemann integrable on $[c, d]$, for the above $\delta > 0$, there exists an $\eta > 0$ such that for any division T of $[c, d]$ with norm $|T| < \eta$, the following relation always holds:

$$\sum_{\alpha} \omega_\alpha \Delta x_\alpha < \frac{\varepsilon\delta}{4M}, \tag{2}$$

where Δx_α is the length of the interval I_α in the division T and

$$\omega_\alpha = \max_{x, y \in I_\alpha} \{|g(x) - g(y)|\}$$

is the oscillation of g on I_α. We recall that the norm $|T|$ is the maximum length of the intervals in T.

Now we consider the composition $f \circ g$. For the division T, let M_α be the oscillation of $f \circ g$ on I_α. Divide all the intervals of the division T into two parts. The first part contains all the intervals on which the oscillation of g is not less than δ, and the second part contains the rest of the intervals. Then we have

$$\sum_{\alpha} M_\alpha \Delta x_\alpha = \sum_{\omega_j \geq \delta} M_j \Delta x_j + \sum_{\omega_i < \delta} M_i \Delta x_i. \tag{3}$$

From (1), we know that for any interval I_i in the second part, $M_i < \varepsilon/2(d - c)$. Thus

$$\sum_{\omega_i < \delta} M_i \Delta x_i < \left(\sum_{\omega_i < \delta} \Delta x_i \right) \cdot \frac{\varepsilon}{2(d-c)} \leq \frac{\varepsilon}{2}, \tag{4}$$

but

$$\sum_{\alpha} \omega_\alpha \Delta x_\alpha \geq \sum_{\omega_j \geq \delta} \omega_j \Delta x_j > \delta \sum_{\omega_j \geq \delta} \Delta x_j. \tag{5}$$

Combining (5) with (2), we obtain

$$\sum_{\omega_j \geq \delta} \Delta x_j < \frac{\varepsilon}{4M}.$$

Then

$$\sum_{\omega_j \geq \delta} M_j \Delta x_j < 2M \cdot \sum_{\omega_j \geq \delta} \Delta x_j < 2M \cdot \frac{\varepsilon}{4M} = \frac{\varepsilon}{2}. \tag{6}$$

Combining (3) with (4) and (6), we have

$$\sum_{\alpha} M_\alpha \Delta x_\alpha < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

That is to say, $f \circ g$ is Riemann integrable on $[c, d]$.

Thus we have proved the following result, which can also be found in [**1**, p. 197].

Proposition 1. *If f is continuous on $[a, b]$ and g is Riemann integrable on $[c, d]$ with its range in $[a, b]$, then $f \circ g$ is Riemann integrable on $[c, d]$.*

Question 2. If f is Riemann integrable on $[a, b]$ and g is continuous on $[c, d]$, is $f \circ g$ always Riemann integrable on $[c, d]$?

The answer is negative, as shown by the following counterexample. Let

$$f(y) = \begin{cases} 0 & \text{when } y = 0, \\ 1 & \text{when } y \neq 0, \end{cases}$$

on $[a, b] = [0, 1]$, and define g inductively as follows.

First, let $g_0(x) = 0$, $x \in [0, 1]$. Next, construct g_1 based on g_0. Divide $[0, 1]$ into three sections, say, I_1, I_2, I_3 in proper order, such that the centre of I_2 is $\frac{1}{2}$ and the length of I_2 is $\frac{1}{3}$. Modifying the function g_0 on I_2 appropriately, we obtain a function g_1, that satisfies the following conditions:

- $g_1(x) = g_0(x)$ for x in I_1 and I_3;
- g_1 is continuous on $[0, 1]$;
- $g_1(x)$ is always greater than zero for any x in the interior of I_2;
- the maximum value of g_1 on I_2 is $\frac{1}{2}$.

Once g_{n-1} is defined, we construct g_n as follows. First, divide all the intervals on which g_{n-1} is always zero into three sections, such that the centre of the middle section is the centre of the original interval and the length of the middle section is $1/3^n \cdot 2^{n-1}$. Second, modify the values of g_{n-1} only on the middle sections of them and obtain a function g_n, such that g_n is still continuous on $[0, 1]$, but in the interior of each modified intervals, g_n is always greater than zero and the maximum is 2^{-n}. We note that there are 2^{n-1} intervals in which g_n and g_{n-1} have different values. Thus the total length of them is 3^{-n}.

Continuing this process gives a sequence of functions $\{g_n\}$ that satisfy the following conditions:

- g_n is continuous on $[0, 1]$;
- $|g_n(x) - g_{n-1}(x)| \leq \frac{1}{2^n}$, for any $x \in [0, 1]$;
- the total length of all the intervals in which g_n is not zero is

$$S_n = \frac{1}{3} + \frac{1}{3^2} + \cdots + \frac{1}{3^n} = \frac{1}{2}\left(1 - \frac{1}{3^n}\right).$$

Thus, for any positive integers $n > m$ we have

$$\begin{aligned} |g_n(x) - g_m(x)| &\leq |g_n(x) - g_{n-1}(x)| + \cdots + |g_{m+1}(x) - g_m(x)| \\ &\leq \frac{1}{2^n} + \cdots \frac{1}{2^{m+1}} < \frac{1}{2^m}. \end{aligned}$$

For any $\varepsilon > 0$, there is a positive integer N, say $N > \ln \varepsilon^{-1}/\ln 2$ when $\varepsilon < 1$. Then for any integers $n > m > N$, we have $|g_n(x) - g_m(x)| < 2^{-N} < \varepsilon$ for any $x \in [0, 1]$. That is to say, $g_n(x)$ is uniformly convergent on $[0, 1]$. Let $g_n(x)$ be uniformly convergent to $g(x)$ on $[0, 1]$. Then g satisfies:

- g is continuous on $[0, 1]$;
- $g(x)$ is not identically zero on any subinterval of $[0, 1]$;
- the total length of all the intervals in which $g(x)$ is not zero is

$$S = \lim_{n \to \infty} \frac{1}{2}\left(1 - \left(\frac{1}{3}\right)^n\right) = \frac{1}{2}.$$

We now prove that $f \circ g$ is not Riemann integrable on $[0, 1]$.

Let T be a division of $[0, 1]$. Divide T into two parts. The first part T_1 contains all the intervals in which $g(x)$ is non-zero and the second part T_2 contains the rest. The total length of all the intervals in T_1 is at most $\frac{1}{2}$; hence the total length of all the intervals in T_2 is at least $\frac{1}{2}$. But in any interval I_i of T_2, we can always find two points ξ_i and ζ_i such that $g(\xi_i) = 0$ and $g(\zeta_i) \neq 0$. Obviously, $f \circ g(\xi_i) = 0$ and $f \circ g(\zeta_i) = 1$. Thus the oscillation M_i of $f \circ g$ on I_i is 1.

Let M_α be the oscillation of $f \circ g$ on any interval I_α of T, and Δx_α be the length of the interval I_α. Then

$$\sum_\alpha M_\alpha \Delta x_\alpha = \sum_{T_1} M_j \Delta x_j + \sum_{T_2} M_i \Delta x_i \geq \sum_{T_2} M_i \Delta x_i = \sum_{T_2} \Delta x_i \geq \frac{1}{2}.$$

Thus $f \circ g$ is not Riemann integrable on $[0, 1]$.

The discussion can be continued by asking for conditions on g to ensure that $f \circ g$ is Riemann integrable, provided that f is Riemann integrable. The following result provides one answer to this question. The proof is left to the reader.

Proposition 2. *Let f be a Riemann integrable function defined on $[a, b]$ and let g be a differentiable function with continuous and non-zero derivative on $[c, d]$. If the range of g is contained in $[a, b]$, then $f \circ g$ is Riemann integrable on $[c, d]$.*

REFERENCE

1. Jonathan Lewin and Myrtle Lewin, *An Introduction to Mathematical Analysis*, Random House, New York, 1988.

Division of Mathematics, School of science, National Institute of Education, Nanyang Technological University, Singapore, 259756.
LUJITAN@HOTMAIL.COM

On the Generalized "Lanczos' Generalized Derivative"

Jianhong Shen

This short note is an extrapolation of Groetsch's interesting article [**1**], and may lead to a clearer understanding of Lanczos' derivative. Only a minimal familiarity with random variables is required.

Lanczos' generalized derivative is defined by

$$D_h f(x) = \frac{3}{2h^3} \int_{-h}^{h} t f(x + t)\, dt,$$

where h is a parameter that can be assumed positive. It generalizes the ordinary derivative in the following two senses:

(1) Suppose $f(x)$ is locally C^4 at x_0. Then $D_h f(x_0) = f'(x_0) + O(h^2)$.

(2) Suppose $f(x)$ has both the right and left derivatives $f'_R(x)$ and $f'_L(x)$ at x_0. Then

$$\lim_{h\to 0} D_h f(x_0) = \frac{f'_R(x_0) + f'_L(x_0)}{2}. \tag{1}$$

A few things puzzled me as I read [**1**]. First, what does the coefficient $(3/2h^3)$ in the definition really mean? Second, how can one see easily from its integral definition that D_h is like a derivative? And finally, how exactly are the right and left derivatives involved in the limiting process of (1)? These questions gave rise to this note.

Let X be a bounded symmetric continuous random variable (i.e., X and $-X$ have the same distribution function) with variance 1. For example, X might be uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$ (with mean 0 and variance 1).

Recall that the ordinary finite difference operator d_h is defined by

$$d_h f(x) = \frac{f(x+h) - f(x)}{h}.$$

For any positive number σ, define

$$L_\sigma f(x) = E\{X^2 d_{\sigma X} f(x)\},$$

where E is the expectation operator.

The motivation is simple. If σ is very small, $Y = \sigma X$ behaves like an atomic distribution at the origin. Therefore, one can pretend that X and Y are independent:

$$L_\sigma f(x) \simeq E\{X^2\} E\{d_Y f(x)\} = E\{d_Y f(x)\}.$$

This is an averaged d_h! Hence, L_σ does resemble the ordinary derivative for small σ.

Moreover, L_σ generalizes Lanczos' derivative D_h. To see this, take X to be any random variable that is uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$. Define $h = \sqrt{3}\sigma$. We show that $L_\sigma = D_h$:

$$\begin{aligned} L_\sigma f(x) &= E\left\{\frac{X}{\sigma}[f(x+\sigma X) - f(x)]\right\} = \frac{1}{\sigma} E\{X f(x+\sigma X)\} \\ &= \frac{1}{\sigma}\int_{-\sqrt{3}}^{\sqrt{3}} t f(x+\sigma t)\frac{dt}{2\sqrt{3}} = \frac{1}{2h}\int_{-\sqrt{3}}^{\sqrt{3}} t f\left(x + \frac{h}{\sqrt{3}}t\right) dt \\ &= \frac{3}{2h^3}\int_{-h}^{h} s f(x+s)\, ds = D_h f(x). \end{aligned}$$

We now understand that the mysterious coefficient $3/2h^3$ has evolved from the simple parameter σ after such a long journey!

A rigorous error estimation for $L_\sigma f(x)$ follows. If $f(x)$ is C^3 near x_0, then

$$d_{\sigma X} f(x_0) = f'(x_0) + \frac{f''(x_0)}{2}\sigma X + O(\sigma^2) \quad \text{as} \quad \sigma \to 0.$$

The error term bound does not depend on the samples of X since we have assumed that X is bounded. Therefore,

$$L_\sigma f(x_0) = E\left\{X^2 f'(x_0) + \frac{f''(x_0)}{2}\sigma X^3 + X^2 O(\sigma^2)\right\} = f'(x_0) + O(\sigma^2).$$

Notice that $E\{X^3\} = 0$ since X is symmetric. This extends the first property of Lanczos' derivative.

The second property of Lanczos' derivative generalizes to L_σ in a similar fashion. Assume that both $f'_R(x_0)$ and $f'_L(x_0)$ exist. Then

$$\begin{aligned} L_\sigma f(x_0) &= E\{X^2 d_{\sigma X} f(x_0) : X > 0\} + E\{X^2 d_{\sigma X} f(x_0) : X < 0\} \\ &= E\{X^2 f'_R(x_0) + X^2 o(1) : X > 0\} + E\{X^2 f'_L(x_0) + X^2 o(1) : X < 0\} \\ &= E\{X^2 f'_R(x_0) : X > 0\} + E\{X^2 f'_L(x_0) : X < 0\} + o(1) \\ &= f'_R(x_0) E\{X^2 : X > 0\} + f'_L(x_0) E\{X^2 : X < 0\} + o(1) \\ &= \frac{f'_R(x_0) + f'_L(x_0)}{2} + o(1). \end{aligned}$$

In the last step, we have applied the symmetry condition and $E\{X^2\} = 1$. The roles of f'_R and f'_L are seen clearly from these five lines.

Finally, notice that: (1) If $f(x)$ is Lipschitz continuous at x_0 with L as its Lipschitz constant, then $|L_\sigma f(x_0)| \leq L$; (2) The random variable involved can be replaced by any suitable distribution with a compact support, since we have not used the positivity condition.

REFERENCE

1. C. W. Groetsch, Lanczos' generalized derivative, *Amer. Math. Monthly* **105** (1998) 320–326.

Computational and Applied Mathematics, UCLA, 7354 Math Sciences Building, Los Angeles, CA 90095
jhshen@math.ucla.edu

A Stability Theorem

Walter Rudin

In 1968 I proved a theorem (stated below) about zeros of holomorphic functions in a polydisc [**2**, p. 87] which was later, in [**1**], referred to, much to my surprise, as a "cornerstone" of multivariable stability theory. The authors of [**1**] pointed out, quite correctly, that my proof used quite a bit of homotopy theory, and they proceeded to prove the theorem by a sequence of more elementary steps. The present note contains an even easier proof, which is also much shorter, and which relies only on very simple properties of the index (or winding number) of a plane curve around the origin.

The following notation will be used. $\mathbf{C}$ is the complex plane, $\mathbf{C}^* = \mathbf{C} \setminus \{0\}$ is the set of all nonzero complex numbers, U and $\overline{U}$ are the open and closed unit discs in $\mathbf{C}$, respectively, and T is the unit circle. For $n \geq 1$,

$$\mathbf{C}^n = \mathbf{C} \times \cdots \times \mathbf{C}, \quad U^n = U \times \cdots \times U, \quad T^n = T \times \cdots \times T;$$

each of these cartesian products has n factors. The torus T^n is the so-called *distinguished boundary* of U^n; it is a small (n-dimensional) part of the whole $(2n-1)$-dimensional boundary of the polydisc U^n.

$A(U^n)$ is the class of all continuous $f: \overline{U^n} \to \mathbf{C}$ that are holomorphic in U^n.

If now $\Gamma: [0, 2\pi] \to \mathbf{C}^*$ is continuous and $\Gamma(2\pi) = \Gamma(0)$ (so that $\Gamma([0, 2\pi])$ is a closed curve in $\mathbf{C}^*$) then there exists a continuous real-valued function α on $[0, 2\pi]$ such that

$$\Gamma(\theta) = |\Gamma(\theta)| \exp\{2\pi i \alpha(\theta)\} \quad (0 \le \theta \le 2\pi). \tag{1}$$

Since $\Gamma(2\pi) = \Gamma(0)$, $\alpha(2\pi) - \alpha(0)$ is an integer (positive, negative, or 0). This is the *index* of Γ:

$$\operatorname{Ind} \Gamma = \alpha(2\pi) - \alpha(0). \tag{2}$$

Note that $\operatorname{Ind} \Gamma$ is independent of the particular choice of α.)

We need the following properties of the index.

(I) Suppose $(s, \theta) \to \Gamma_s(\theta)$ is a continuous map from $[0, 1] \times [0, 2\pi]$ into $\mathbf{C}^*$, and $\Gamma_s(2\pi) = \Gamma_s(0)$ for all s. *Then* $\operatorname{Ind} \Gamma_s$ *is the same for all* s.

The reason is simply that $\operatorname{Ind} \Gamma_s$ is a *continuous* function of s. Being integer-valued, this function is constant on the connected set $[0, 1]$.

(II) *If* $G : \overline{U} \to \mathbf{C}^*$ *is continuous and if we define* $G|_T(\theta) = G(e^{i\theta})(0 \le \theta \le 2\pi)$ *then* $\operatorname{Ind} G|_T = 0$.

To deduce this from (I) put $\Gamma_s(\theta) = G(se^{i\theta})$ and note that $\Gamma_1 = G|_T$, Γ_0 is the constant $G(0)$.

(III) *If* $h \in A(U)$ *and* $h(T) \subset \mathbf{C}^*$ *then* $\operatorname{Ind} h|_\Gamma$ *is equal to the number of zeros of* h *in* U.

This is the classical "argument principle" of complex analysis.

Theorem. *Suppose* $\Phi = (\varphi_1, \ldots, \varphi_n)$ *is a continuous map of* $\overline{U}$ *into* $\overline{U^n}$ *that carries* T *into* T^n, *such that*

$$\operatorname{Ind} \varphi_j|_T > 0 \quad \textit{for} \quad 1 \le j \le n \tag{3}$$

Put $K = \Phi(\overline{U})$. *Then*

$$f(T^n \cup K) = f(\overline{U^n}) \tag{4}$$

for every $f \in A(U^n)$.

Proof: Assume $f(z) \neq 0$ for every $z \in T^n \cup K$. We show that $f(z) \neq 0$ for every $z \in \overline{U^n}$. This implies the theorem, and shows why the term "stability" was used in this connection.

Fix $a = (a_1, \ldots, a_n) \in \overline{U^n}$. Let $\operatorname{Ind} \varphi_j|_T = m_j$. There exist $c_j \in \mathbf{C}$ such that

$$c_j^{m_j} = a_j \quad (1 \le j \le n). \tag{5}$$

Since $m_j > 0$, $|c_j| \le 1$. Define

$$h(\lambda) = f\left(\left(\frac{\lambda + c_1}{1 + \overline{c_1}\lambda}\right)^{m_1}, \ldots, \left(\frac{\lambda + c_n}{1 + \overline{c_n}\lambda}\right)^{m_n}\right) \tag{6}$$

for $\lambda \in \overline{U}$. Then $h \in A(U)$, $h(T) \subset \mathbf{C}^*$, $h(0) = f(a)$. Hence $f(a) \neq 0$ follows from

$$\operatorname{Ind} h|_T = 0 \tag{7}$$

because of (III).

Since $(f \circ \Phi)(\overline{U}) = f(K) \subset \mathbf{C}^*$, by assumption, (II) shows that

$$\operatorname{Ind} f \circ \Phi|_T = 0. \tag{8}$$

There are continuous real-valued functions α_j, β_j such that

$$\varphi_j(e^{i\theta}) = \exp\{2\pi i\alpha_j(\theta)\}, \quad \left(\frac{e^{i\theta}+c_j}{1+\bar{c}_j e^{i\theta}}\right)^{m_j} = \exp\{2\pi i\beta_j(\theta)\}$$

on $[0, 2\pi]$. Note that

$$\alpha_j(2\pi) - \alpha_j(0) = m_j = \beta_j(2\pi) - \beta_j(0). \tag{9}$$

Define

$$\gamma_{j,s}(\theta) = s\alpha_j(\theta) + (1-s)\beta_j(\theta) \quad (0 \le s \le 1, 0 \le \theta \le 2\pi) \tag{10}$$

and let $\Psi_s\colon [0, 2\pi] \to T^n$ be the map whose j^{th} component is $\exp\{2\pi i\gamma_{j,s}(\theta)\}$. Then

$$\Psi_s(2\pi) = \Psi_s(0) \quad (0 \le s \le 1), \tag{11}$$

$\Psi_s(T) \subset T^n$, hence $f(\Psi_s(T)) \subset \mathbf{C}^*$, and now (I) shows that

$$\operatorname{Ind} f \circ \Psi_1 = \operatorname{Ind} f \circ \Psi_0. \tag{12}$$

Since $f \circ \Phi|_T = f \circ \Psi_1$ and $h|_T = f \circ \Psi_0$, (12) and (8) imply (7). ∎

Remarks. (i) The simplest example of a Φ as in the theorem is $\Phi(\lambda) = (\lambda, \lambda, \ldots, \lambda)$. Then K is a disc (2-dimensional), $\dim(T^n \cup K) = n$, whereas $\dim U^n = 2n$.

(ii) It is not necessary for Φ to map U into the interior U^n of $\overline{U^n}$. For example, when $n = 2$,

$$\Phi(\mathrm{re}^{i\theta}) = \begin{cases} (2re^{i\theta}, 0) & (0 \le r \le 1/2) \\ (e^{i\theta}, 2r-1) & (1/2 \le r \le 1) \end{cases}$$

will do nicely.

(iii) The hypothesis "$m_j > 0$ for all j" cannot be omitted. To see this, take $n = 2$, $\Phi(\lambda) = (\lambda, \bar{\lambda})$. Then $m_1 = 1$, $m_2 = -1$. If $f(z, w) = 1 + 4zw$, then $|f| \ge 1$ on $T^2 \cup \Phi(\overline{U})$ but $f(\frac{1}{2}, -\frac{1}{2}) = 0$.

For another example, take $\Phi(\lambda) = (\lambda, 1)$, so that $m_1 = 1$, $m_2 = 0$, and put $f(z, w) = 2w - 1$. Then $|f| \ge 1$ on $T^2 \cup \Phi(\overline{U})$ but $f(z, 1/2) = 0$ for all z.

However, the hypothesis "$m_j > 0$ for all j" *can* be replaced by "$m_j < 0$ for all j" because the theorem can then be applied to $\Phi(\bar{\lambda})$ in place of $\Phi(\lambda)$.

REFERENCES

1. Ph. Delsarte, Y. Genin, Y. Kamp, A simple proof of Rudin's multivariable stability theorem, Philips Research Laboratory Report R412, Brussels, Dec. 1979.
2. W. Rudin, *Function Theory in Polydiscs*, Benjamin, 1969.

University of Wisconsin-Madison, Madison, Wisconsin 53706
recep@math.wisc.edu

Rationals and the Modular Group

Roger C. Alperin

The modular group $\mathcal{M}$ is the quotient group $PSL_2(\mathbf{Z}) = SL_2(\mathbf{Z})/\{\pm I\}$ of $SL_2(\mathbf{Z})$, the group of 2×2 integer matrices of determinant 1. In [1] we gave an elementary proof that $\mathcal{M}$ has the structure of a free product of a cyclic group of order 2 generated by the image of $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and a cyclic group of order 3 generated by the image of $B = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$.

The free product structure provides a description of the non-trivial elements of $\mathcal{M}$ as unique strings of A's and B's with the property that there are no two consecutive A's and no three consecutive B's; we refer to these as *reduced strings*. We explained this free product structure in terms of the action of the modular group on the irrationals. In this note we describe the action on the rationals; this can be viewed as a way of describing the inverse of the Euclidean algorithm.

The group $SL_2(\mathbf{Z})$ acts via linear transformations on $\mathbf{R}^2$ as column vectors and this gives an action of $\mathcal{M}$ via linear fractional transformations on the projective line $P^1(\mathbf{R})$, the real numbers together with ∞. We may also view $P^1(\mathbf{R})$ as the slopes of non-zero vectors, that is, the equivalence classes of $\mathbf{R}^2 - \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ induced by non-zero scalar multiplication; the equivalence class of the vector $e = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is denoted ∞, the equivalence class of the vector $\begin{pmatrix} p \\ q \end{pmatrix}$, $q \neq 0$ is the same as that of $\begin{pmatrix} p/q \\ 1 \end{pmatrix}$ and corresponds to the real number $z = p/q$. For the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ in $SL_2(\mathbf{Z})$ the induced action on $P^1(\mathbf{R})$ is given by

$$z \longrightarrow \frac{az + b}{cz + d}.$$

The induced action for the generating elements is given as

$$A: z \longrightarrow \frac{-1}{z}, \quad B: z \longrightarrow \frac{-1}{z+1}, \quad B^2: z \longrightarrow -1 - \frac{1}{z}.$$

The orbit of e is easily seen to be in correspondence with the set of all first columns of matrices from $SL_2(\mathbf{Z})$. Thus the orbit of ∞ is in 1-1 correspondence with the projective line $P^1(\mathbf{Q})$, consisting of the set $\mathbf{Q}$ of all reduced fractions together with ∞; from elementary group theory this is in 1-1 correspondence with the set of left cosets of the stabilizer $\mathcal{N}$ of e, which is the image of the subgroup generated by AB in $\mathcal{M}$.

Using the free product description of $\mathcal{M}$ we can also describe the set of coset representatives as reduced strings of A's and B's . First, a non-trivial coset representative cannot end in AB or its inverse B^2A; therefore if it ends in A it is either A or of the form ZBA with Z ending in A or trivial; if it ends in B it is either B or ZB^2 with Z ending in A or trivial. Thus, as a first pass, the set of coset representatives is the set $\mathcal{R} = \{I\} \cup \{A\} \cup \{B\} \cup \{BA\} \cup \{B^2\} \cup \{ZBA | Z$ any string ending in $A\} \cup \{ZB^2 | Z$ any string ending in $A\}$. Next, to determine the distinct coset representatives we just observe that the free product description

gives a unique expression for the elements. The coset equivalence relation $X \cong Y$ on reduced strings $X, Y \in \mathscr{M}$ is $Y = X(AB)^n$ or $Y = X(B^2A)^n$ for some non-negative integer n. We see easily that $A \cong B \bmod \mathscr{N}$ and hence also $XA \cong XB \bmod \mathscr{N}$ for any string X, and hence if this is reduced, $X \neq I$ must end in B. Thus we can simplify the description of the distinct coset representatives to $\mathscr{R} = \{I\} \cup \{B\} \cup \{B^2\} \cup \{ZB^2 | Z \text{ any string ending in } A\}$. It is easy to see that no two of these reduced strings are equivalent; for example, for $Z \neq W$, both ending in A, then $ZB^2 = WB^2(B^2A)^n$ and $ZB^2 = WB^2(AB)^n$ are impossible. Thus the coset representatives of $\mathscr{M}/\mathscr{N}$ are the distinct strings $\mathscr{R} = \{I\} \cup \{B\} \cup \{B^2\} \cup \{ZB^2 | Z$ any string ending in $A\}$.

We can also describe this set $\mathscr{R}$ as the union of $\mathscr{R}_m$ defined inductively as

$$\begin{aligned} \mathscr{R}_0 &= \{I, B\}, \quad \mathscr{N}_0 = \{B^2\} \\ \mathscr{P}_m &= \{A\}\mathscr{N}_m, \quad \mathscr{N}_{m+1} = \{B^2, B\}\mathscr{P}_m \qquad (1) \\ \mathscr{R}_{m+1} &= \mathscr{R}_m \cup \mathscr{P}_m \cup \mathscr{N}_m \end{aligned}$$

It is easy to see that $\mathscr{R}_m$ has 2^{m+1} elements and $\mathscr{P}_m$, $\mathscr{N}_m$ each have 2^m elements. We can rewrite (1) as

$$\mathscr{P}_{m+1} = \{AB, AB^2\}\mathscr{P}_m, \quad \mathscr{N}_{m+1} = \{B^2A, BA\}\mathscr{N}_m. \qquad (2)$$

Simplifying (2) we obtain the following result. $\mathscr{FS}(x, y)$ denotes the free semigroup with the generators, x, y.

Proposition. *$P^1(\mathbf{Q})$ is in 1-1 correspondence with $\mathscr{R} = \{I\} \cup \{B\} \cup \mathscr{FS}(AB, AB^2) \cdot AB^2 \cup \mathscr{FS}(B^2A, BA) \cdot B^2$.*

Observing, that $I\infty = \infty$, $B\infty = 0$, $B^2\infty = -1$ and $AB^2\infty = 1$, and $P^1(\mathbf{Q}) - \{\infty, 0\}$ is the positive and negative rationals, we have the following

Corollary. *The set of positve rationals is the orbit of the free semigroup generated by AB and AB^2 on $z = 1$. The set of negative rationals is the orbit of the free semigroup generated by B^2A and BA on $z = -1$.*

These upper $U = AB$, $U^- = B^2A$ and lower $L = AB^2$, $L^- = BA$ triangular matrix actions corresponding to these semigroup generators are

$$U: z \to z + 1, \quad L: z \to \frac{z}{z+1},$$

$$U^-: z \to z - 1, \; L^-: z \to \frac{z}{1-z}.$$

Every positive rational is uniquely expressible in terms of semigroup generators as an element of the orbit of 1. Alternatively, starting from a reduced positive

rational $z = \frac{p}{q}$ we can apply a greedy or Euclidean recursion to obtain a finite sequence that stabilizes at 1;

$$\frac{p}{q} \to \begin{cases} \dfrac{(p-q)}{q} & \text{if } p > q \\ \dfrac{p}{(q-p)} & \text{if } p < q \\ 1 & \text{if } p = q = 1 \end{cases}$$

Here we apply U^- or L^- depending on whether or not $z > 1$ or $z < 1$. For example, the sequence

$$\frac{34}{55}, \frac{34}{21}, \frac{13}{21}, \frac{13}{8}, \frac{5}{8}, \frac{5}{3}, \frac{2}{3}, \frac{2}{1}, 1$$

corresponds to $(U^-L^-)^4(\frac{34}{55}) = 1$. Since we have shown that the set of positive rationals can be described by a free semigroup, this means that $(LU)^4L$ is *the* coset representative in $\mathscr{R}_9$ corresponding to $34/55$ as described in the Corollary.

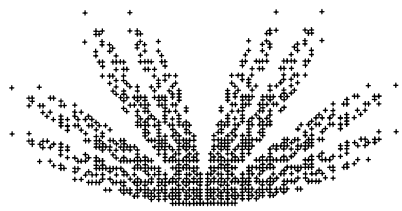


Figure 1. $\mathscr{R}_9$.

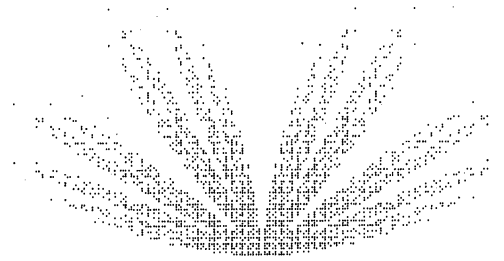


Figure 2. $\mathscr{R}_{10}$.

Finally, we consider the matrix action of the distinct non-trivial coset representatives in $\mathscr{R}$ on the column $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and the plots of these images in $\mathbf{R}^2$. These images are just the points with relatively prime coordinates in the upper half-plane. We obtain the fascinating plant-like structures in Figures 1 and 2. For example, the distant points from the root $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ are the 'Fibonacci points' $\begin{pmatrix} \pm 34 \\ 55 \end{pmatrix}$, $\begin{pmatrix} \pm 55 \\ 34 \end{pmatrix}$ in $\mathscr{R}_9$.

REFERENCE

1. Roger C. Alperin, $PSL_2(\mathbf{Z}) = \mathbf{Z}_2 * \mathbf{Z}_3$ *Amer. Math. Monthly* **100** (1993) 385–386.

alperin@mathcs.sjsu.edu

The Union of Vieta's and Wallis's Products for Pi

Thomas J. Osler

The beautiful infinite product of radicals

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}\cdots \tag{1}$$

due to Vieta in 1592 [**2**], is one of the oldest noniterative analytical expressions for π. Wallis's product dating from 1655 [**3**]

$$\frac{2}{\pi} = \frac{1\cdot 3}{2\cdot 2}\cdot\frac{3\cdot 5}{4\cdot 4}\cdot\frac{5\cdot 7}{6\cdot 6}\cdot\frac{7\cdot 9}{8\cdot 8}\cdots \tag{2}$$

is also most remarkable. Both are usually included in any list of interesting expressions for π [**1**].

The purpose of this short note is to call attention to the following union of Vieta and Wallis-like products:

$$\frac{2}{\pi} = \prod_{n=1}^{p}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\cdots+\frac{1}{2}\sqrt{\frac{1}{2}}}}}$$

$$(n \text{ radicals})$$

$$\times\prod_{n=1}^{\infty}\frac{2^{p+1}n-1}{2^{p+1}n}\cdot\frac{2^{p+1}n+1}{2^{p+1}n}. \tag{3}$$

While (1) and (2) seem unrelated, they are both special cases of a more general double product (3). The first product in (3) consists of the first p factors of Vieta's original infinite product (1). The second product in (3) is a Wallis-like product. We say this because the case $p = 0$ gives us Wallis's original product (2), and for other values of p it is Wallis's product with factors deleted. Notice also that the Wallis-like product in (3) provides us with the error factor needed to make the Vieta product (1) exact when only finitely many factors are used.

Relation (3) yields Vieta's product (1) when p goes to infinity, and Wallis's product (2) when $p = 0$. For each intermediate value of $p = 1, 2, 3, \ldots$ we obtain

united Vieta-Wallis-like products:

$$p = 0: \quad \frac{2}{\pi} = \frac{1\cdot 3}{2\cdot 2}\cdot\frac{3\cdot 5}{4\cdot 4}\cdot\frac{5\cdot 7}{6\cdot 6}\cdot\frac{7\cdot 9}{8\cdot 8}\cdot\frac{9\cdot 11}{10\cdot 10}\frac{11\cdot 13}{12\cdot 12}\cdots$$

(Wallis's original product)

$$p = 1: \quad \frac{2}{\pi} = \sqrt{\frac{1}{2}}\cdot\frac{3\cdot 5}{4\cdot 4}\cdot\frac{7\cdot 9}{8\cdot 8}\cdot\frac{11\cdot 13}{12\cdot 12}\cdot\frac{15\cdot 17}{16\cdot 16}\cdot\frac{19\cdot 21}{20\cdot 20}\cdots$$

$$p = 2 \quad \frac{2}{\pi} = \sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\cdot\frac{7\cdot 9}{8\cdot 8}\cdot\frac{15\cdot 17}{16\cdot 16}\cdot\frac{23\cdot 25}{24\cdot 24}\cdot\frac{31\cdot 33}{32\cdot 32}\cdots$$

$$p = 3: \quad \frac{2}{\pi} = \sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}\cdot\frac{15\cdot 17}{16\cdot 16}\cdot\frac{31\cdot 33}{32\cdot 32}\cdot\frac{47\cdot 49}{48\cdot 48}\cdot\frac{63\cdot 65}{64\cdot 64}\cdots$$

$\cdots$

$$p \to \infty: \quad \frac{2}{\pi} = \sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}\cdots$$

(Vieta's original product)

An examination of these special cases of (3) shows that each time we increase p by one, we insert one new radical factor in the Vieta-like product, and remove alternate factors from the Wallis-like product. The author accidently discovered (3) while trying to derive (1).

To derive (3) we start by applying the double angle formula for the sine function p times to obtain

$$\begin{aligned}
\sin\theta &= 2\cos\frac{\theta}{2}\sin\frac{\theta}{2}\\
&= 2^2\cos\frac{\theta}{2}\cos\frac{\theta}{2^2}\sin\frac{\theta}{2^2}\\
&= 2^3\cos\frac{\theta}{2}\cos\frac{\theta}{2^2}\cos\frac{\theta}{2^3}\sin\frac{\theta}{2^3}\\
&\ \ \vdots\\
\sin\theta &= 2^p\cos\frac{\theta}{2}\cos\frac{\theta}{2^2}\cos\frac{\theta}{2^3}\cdots\cos\frac{\theta}{2^p}\sin\frac{\theta}{2^p} \qquad (4)
\end{aligned}$$

Next we use the infinite product for the sine function [4], valid for all x,

$$\sin x = x\prod_{n=1}^{\infty}\left(1-\frac{x^2}{\pi^2 n^2}\right) = x\prod_{n=1}^{\infty}\left(\frac{\pi n - x}{\pi n}\cdot\frac{xn + x}{\pi n}\right)$$

with $x = \theta/2^p$ to replace the last factor in 4. Dividing by θ gives

$$\frac{\sin\theta}{\theta} = \cos\frac{\theta}{2}\cos\frac{\theta}{2^2}\cos\frac{\theta}{2^3}\cdots\cos\frac{\theta}{2^p}\prod_{n=1}^{\infty}\left(\frac{2^p\pi n - \theta}{2^p\pi n}\cdot\frac{2^p\pi n + \theta}{2^p\pi n}\right). \qquad (5)$$

Now express each of the cosine factors in (5) in terms of $\cos\theta$ by repeated use of the half-angle formula for the cosine; here we assume $-\pi/2 \le \theta \le \pi/2$ so that the cosines are never negative.

$$\cos\frac{\theta}{2} = \sqrt{\frac{1}{2} + \frac{1}{2}\cos\theta}$$

$$\cos\frac{\theta}{2^2} = \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\cos\theta}}$$

$$\vdots$$

$$\cos\frac{\theta}{2^p} = \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \cdots + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\cos\theta}}}} \qquad (6)$$

(p *radicals*)

Combining (6) with (5) we obtain

$$\frac{\sin\theta}{\theta} = \prod_{n=1}^{p} \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \cdots + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\cos\theta}}}}$$

(n *radicals*)

$$\times \prod_{n=1}^{\infty} \left(\frac{2^p\pi n - \theta}{2^p\pi n} \cdot \frac{2^p\pi n + \theta}{2^p\pi n} \right) \qquad (7)$$

If we set $\theta = \pi/2$ in (7) and simplify we obtain (3). ∎

REFERENCES

1. L. Berggren, J. Borwein, and P. Borwein, *Pi, A Source Book*, Springer, New York, 1997, pp. 686–689.
2. F. Vieta, Variorum de Rebus Mathematicis Reponsorum Liber VII, (1593) in: *Opera Mathematica*, (*reprinted*) Georg Olms Verlag, Hildesheim, New York, 1970, pp. 398–400 and 436–446.
3. J. Wallis, Computation of π by Successive Interpolations, (1655) in: *A Source Book in Mathematics, 1200–1800* (D. J. Struik, Ed.), Harvard University Press, Cambridge, MA, 1969, pp. 244–253.
4. E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, Fourth Ed., 1927, p. 137.

Rowan University, Glassboro, NJ 08028
osler@rowan.edu

PROBLEMS AND SOLUTIONS

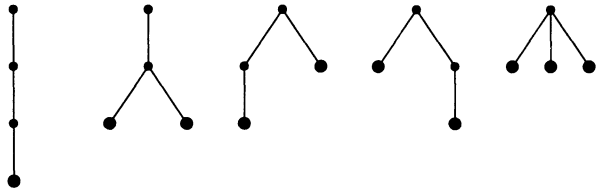
Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Mario Benedicty, Paul Bracken, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Glenn G. Chappell, Randall Dougherty, Roger B. Eggleton, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Kiran S. Kedlaya, Murray S. Klamkin, Fred Kochman, Frederick W. Luttmann, Vania Mascioni, Frank B. Miles, Richard Pfiefer, Cecil C. Rousseau, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before March 31, 2000; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

PROBLEMS

10753. *Proposed by Louis Shapiro, Howard University, Washington, DC.* An *ordered tree* is a rooted tree in which the children of each node form a sequence as opposed to a set. The 5 ordered trees with 3 edges are



The number of ordered trees with n edges is the nth Catalan number $\binom{2n}{n}/(n+1)$. Therefore, if one draws each of the ordered trees with n edges, one draws a total of $\binom{2n}{n}$ nodes. Prove that exactly half of these nodes are end-nodes (i.e., leaves with no children).

10754. *Proposed by Paul Bracken, Université de Montréal, Montréal, PQ, Canada.* Let $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$, and let $\rho(s,n) = \sum_{k=n+1}^{\infty} k^{-s}$. Show that for positive integers $s \geq 2$,

$$\sum_{k=1}^{\infty} \frac{\rho(s,k)}{k} = \frac{s}{2}\zeta(s+1) - \frac{1}{2}\sum_{k=1}^{s-2} \zeta(s-k)\zeta(k+1).$$

10755. *Proposed by Jiro Fukuta, Motosu-gun, Gifu-ken, Japan.* An arbitrary circle O is drawn through vertices B and D of a convex quadrilateral $ABCD$. Let O_1 be the circle tangent to lines AB and AD and tangent to O internally at a point of O on the opposite side of line BD from A. Let O_2 be the circle tangent to lines CB and CD and tangent to O internally at a point of O on the opposite side of line BD from C. Let R_1 and R_2 be the radii of circles O_1 and O_2, respectively, and let r_1 and r_2 be the radii of the incircles of triangles ABD and CBD, respectively. Prove that the quadrilateral $ABCD$ is inscribable in a circle if and only if $r_1/R_1 + r_2/R_2 = 1$.

10756. *Proposed by Douglas Iannucci, University of the Virgin Islands, St. Thomas, VI.* Prove that

$$\cos\frac{\pi}{7} = \frac{1}{6} + \frac{\sqrt{7}}{6}\left(\cos\left(\frac{1}{3}\arccos\frac{1}{2\sqrt{7}}\right) + \sqrt{3}\sin\left(\frac{1}{3}\arccos\frac{1}{2\sqrt{7}}\right)\right).$$

10757. *Proposed by Mark Kidwell, United States Naval Academy, Annapolis, MD.* Given integers $a_0, a_1, a_2, \ldots, a_n$ with $a_i \neq 0$ for $i \geq 1$, write $[a_0; a_1, a_2, \ldots, a_n]$ for the continued fraction

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \frac{1}{a_n}}}}.$$

Every positive rational number has a unique representation as $[a_0; a_1, a_2, \ldots, a_n]$ if we require that $a_0 \geq 0$, $a_i > 0$ for $1 \leq i \leq n-1$, and $a_n > 1$ (we call this the *standard representation*), but it can have other representations $[b_0; b_1, b_2, \ldots, b_m]$ if we permit negative values for some of the b_i or if we permit $b_m = 1$. For example, $11/3 = [3; 1, 2] = [3; 1, 1, 1] = [4; -3]$. Prove or disprove: If r is a positive rational number, $r = [a_0; a_1, a_2, \ldots, a_n]$ is the standard representation, and $r = [b_0; b_1, b_2, \ldots, b_m]$ is another representation, then $a_0 + a_1 + \cdots + a_n \leq |b_0| + |b_1| + \cdots + |b_m|$, with strict inequality if any of the b_i are negative.

10758. *Proposed by Mark Sapir, Vanderbilt University, Nashville, TN.* Prove that the sum of the (decimal) digits of 9^n cannot equal 9 when $n > 2$.

10759. *Proposed by Călin Popescu, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.* In triangle ABC, let h_a denote the altitude to the side BC and let r_a denote the exradius relative to side BC, i.e., the radius of the circle tangent to the extensions of sides AB and AC and to the side BC externally. Define h_b, h_c, r_b, and r_c correspondingly. Prove that $h_a^n r_a^n + h_b^n r_b^n + h_c^n r_c^n \leq r_a^n r_b^n + r_b^n r_c^n + r_c^n r_a^n$ for any integer n, and determine conditions for equality.

SOLUTIONS

Common Eigenvector of Commuting Matrices

10633 [1997, 975]. *Proposed by Kiran S. Kedlaya, Princeton University, Princeton, NJ.* Let S be a commuting family of n-by-n matrices over an arbitrary field. Suppose the matrices in S have a common eigenvector v, so that $Mv = \lambda_M v$ for all $M \in S$. Prove that the transposes of these matrices also have a common eigenvector with these eigenvalues, that is, a vector w satisfying $M^T w = \lambda_M w$ for all $M \in S$.

Solution by Alain Tissier, Montmermeil, France. Let K be the field. Set $\phi(M) = M - \lambda_M I$ and $\phi(S) = \{\phi(M) : M \in S\}$. Thus $\phi(S)$ is a commuting family of $n \times n$ matrices over K having a common nonzero vector v such that $\phi(M)v = 0$ for all $\phi(M) \in \phi(S)$. Since $\phi(M)^T = M^T - \lambda_M I$, we have to prove only that the transposes of the matrices in $\phi(S)$ have a common nonzero vector w satisfying $\phi(M)^T w = 0$ for $\phi(M) \in \phi(S)$. Thus we may suppose that $\lambda_M = 0$ for every M.

If all matrices in S are nilpotent, then the collection of transposes is also a commuting family of nilpotent matrices. In this case there is a nonzero vector w such that $M^T w = 0$ for all $M \in S$ (section 3.3 of J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, 1972). So we may assume that not all elements of S are nilpotent.

We proceed by induction on n. When $n = 1$ all the matrices are zero, so the conclusion is true. Take $n > 1$, and suppose the result is true for h-by-h matrices for each $h < n$. Let N

be a nonnilpotent element of S. Let W be the set of all vectors x such that $N^k x = 0$ for some $k \geq 0$. By finite-dimensionality, there is a fixed k such that $N^k x = 0$ for all $x \in W$. So $v \in W$, W is a subspace, and $K^n = W \oplus U$, where U is the range of the mapping $x \mapsto N^k x$. Now if $M \in S$, then M commutes with N, and the descriptions of W and U show that they are invariant under M. Let m be the dimension of W, let $\mathcal{B}'$ be a basis of W, and let $\mathcal{B}''$ be a basis of U. For each $M \in S$, let M' be the $\mathcal{B}'$-representation of M restricted to W and let M'' be the $\mathcal{B}''$-representation of M restricted to U. Then there exists a nonsingular $n \times n$ matrix P such that $P^{-1}MP = \begin{bmatrix} M' & 0 \\ 0 & M'' \end{bmatrix}$ for all $M \in S$. Let S' be the set of the matrices M'. Then S' is a family of $m \times m$ commuting matrices having a common nonzero vector v' such that $M'v' = 0$ for each $M' \in S'$. By the induction hypothesis there exists a nonzero vector w' such that ${M'}^T w' = 0$ for each $M' \in S'$. The vector $(P^T)^{-1} \begin{bmatrix} w' \\ 0 \end{bmatrix}$ solves the problem.

Solved also by R. J. Chapman (U. K.), D. Huang, J. H. Lindsey II, G. Sansigre Vidal (Spain), GCHQ Problems Group (U. K.), and the proposer.

Reflected Concurrent Lines

10637 [1998, 68]. *Proposed by C. F. Parry, Exmouth, Devon, United Kingdom.* Suppose triangle ABC has circumcircle Γ, circumcenter O, and orthocenter H. Parallel lines α, β, γ are drawn through the vertices A, B, C, respectively. Let α', β', γ' be the reflections of α, β, γ in the sides BC, CA, AB, respectively.

(**a**) Show that α', β', γ' are concurrent if and only if α, β, γ are parallel to the Euler line OH.

(**b**) Suppose that α', β', γ' are concurrent at the point P. Show that Γ bisects OP.

Solution by Robert L. Young, Osterville, MA. Take Γ to be the unit circle $z\overline{z} = 1$ in the complex plane and rotate ABC about O so that $\arg H = 0$. Assume $H \neq 0$ for now, so the Euler line exists and is the real axis. Choose $\theta_3 > \theta_2 > \theta_1 > 0$ so that $A = e^{i\theta_1}$, $B = e^{i\theta_2}$, and $C = e^{i\theta_3}$, and let $M = e^{i\theta}$, where $\theta \in [0, \pi)$ is the angle of inclination of the lines α, β, γ.

(**a**) The reflection z' of a complex number z through the line containing B and C is determined as follows. Apply the linear transformation $\tau(z) = (z - B)\overline{(C - B)}$, which takes B and C and therefore the line BC to the real axis. Since reflection in the real axis is conjugation,

$$z' = \tau^{-1}(\overline{\tau(z)}) = \frac{(\overline{z - B})(C - B)}{(\overline{C} - \overline{B})} \frac{BC}{BC} + B = -BC\overline{z} + B + C,$$

and the reflection of A through line BC is

$$A' = -BC\overline{A} + B + C. \tag{1}$$

Any $z \neq A'$ on α' satisfies the equation

$$\frac{z - A'}{\overline{z} - \overline{A'}} = e^{2i \arg \alpha'}. \tag{2}$$

Since the perpendicular bisector of line BC passes through O and $\exp(i(\theta_2 + \theta_3)/2)$, we have $\arg(C - B) \equiv (\theta_2 + \theta_3)/2 - \pi/2$ modulo π. By the definition of α', $\arg \alpha' + \arg \alpha \equiv 2\arg(C - B) \equiv \theta_2 + \theta_3 - \pi$ modulo 2π, so $e^{2i \arg \alpha'} = e^{i(2\theta_2 + 2\theta_3 - 2\theta)} = B^2C^2\overline{M}^2$. Substituting (1) into (2), we conclude that α' has equation

$$z = \overline{M}^2 C^2 B^2 \left(\overline{z} + A\,\overline{B}\,\overline{C} - \overline{B} - \overline{C}\right) - BC\,\overline{A} + B + C.$$

It is convenient to note that $A + B + C = H$ and is therefore real and to write $K = ABC$, so that $AB + BC + CA = K\overline{H} = KH$. With this notation, the equation becomes $z = \overline{M}^2K^2\overline{A}^2(\overline{z} + (A - C - B)\overline{B}\,\overline{C}) + (AB + AC - BC)\overline{A}$, or

$$z = K\big(\overline{M}^2K\overline{z} - 2\big)\overline{A}^2 - \big(\overline{M}^2 - 1\big)KH\overline{A} + 2\overline{M}^2K.$$

Similarly, the equation of β' is

$$z = K\big(\overline{M}^2K\overline{z} - 2\big)\overline{B}^2 - \big(\overline{M}^2 - 1\big)KH\,\overline{B} + 2\overline{M}^2K.$$

Let z_C denote point of intersection, if any, of α' and β' and similarly for z_A and z_B. Solving for z_C from these two equations, we get $K\big(\overline{M}^2K\overline{z_C} - 2\big)\overline{A}^2 - \big(\overline{M}^2 - 1\big)KH\,\overline{A} = K\big(\overline{M}^2K\overline{z_C} - 2\big)\overline{B}^2 - \big(\overline{M}^2 - 1\big)KH\,\overline{B}$, so $K\big(\overline{A}^2 - \overline{B}^2\big)\big(\overline{M}^2K\overline{z_C} - 2\big) = \big(\overline{A} - \overline{B}\big)\big(\overline{M}^2 - 1\big)KH$, and

$$\big(\overline{M}^2K\overline{z_C} - 2\big)\big(\overline{A} + \overline{B}\big) = \big(\overline{M}^2 - 1\big)H.$$

Similarly,

$$\big(\overline{M}^2K\overline{z_B} - 2\big)\big(\overline{A} + \overline{C}\big) = \big(\overline{M}^2K\overline{z_A} - 2\big)\big(\overline{B} + \overline{C}\big) = \big(\overline{M}^2 - 1\big)H.$$

Suppose α', β', γ' are concurrent at P. Then $\big(\overline{A} + \overline{B}\big)\big(\overline{M}^2K\overline{P} - 2\big)$, $\big(\overline{B} + \overline{C}\big)\big(\overline{M}^2K\overline{P} - 2\big)$, and $\big(\overline{C} + \overline{A}\big)\big(\overline{M}^2K\overline{P} - 2\big)$ all equal $\big(\overline{M}^2 - 1\big)H$. Multiply the first of these equations by $\overline{B} + \overline{C}$, multiply the second by $\overline{A} + \overline{B}$, and then subtract to obtain $0 = \big(\overline{M}^2 - 1\big)H\big(\overline{A} - \overline{C}\big)$. Since $A \neq C$ and $H \neq 0$, we have $M^2 = 1$ and $\theta = 0$. So α, β, γ are parallel to the Euler line as claimed. Conversely, if α, β, γ are parallel to the Euler line, then $M^2 = 1$, and $z_A = z_B = z_C = P = 2K$ satisfy the equations for α', β', γ', so these are concurrent.

If $H = 0$, there is no Euler line. In this case, α', β', and γ' concur at $P = 2K\,\overline{M}^2$.
(b) Since $P = 2K = 2ABC$, we have $|P| = 2$. Therefore $|(O + P)/2| = 1$ and $(O+P)/2$ is on Γ.

Solved also by J. Anglesio (France), M. Benedicty, N. Lakshmanan, and V. Schindler (Germany).

A Constrained Maximization

10646 [1998, 176]. *Proposed by Hassan Ali Shah Ali, Teheran, Iran.* Find the maximum of $\prod_{i=1}^{n}(1 - x_i)$ over all nonnegative $x_1, x_2, \ldots, x_n$ with $\sum_{i=1}^{n} x_i^2 = 1$.

Solution by Patrick A. Staley, Southwestern College, Chula Vista, CA. When $n = 1$, the constraint requires $x_1 = 1$, and the maximum value is 0. So assume $n \geq 2$. We show that the maximum is $3/2 - \sqrt{2} \approx 0.0858$, and it occurs when two of the x_i's are $1/\sqrt{2}$ and the others are 0.

Let $x_1, x_2, \ldots, x_n$ be an optimal solution. If x and y are any two of the x_i's, then they satisfy a two-element subproblem: maximize $(1 - x)(1 - y)$ under the constraints $x \geq 0$, $y \geq 0$, and $x^2 + y^2 = k^2$ for a given positive $k \leq 1$. To solve this, note that $dy/dx = -x/y$, so

$$\frac{d\big((1 - x)(1 - y)\big)}{dx} = -(1 - y) - (1 - x)\frac{dy}{dx} = \frac{(x - y)(1 - x - y)}{y}.$$

If this vanishes, then $(x + y - 1)(x - y) = 0$. There are three possibilities for the global maximum of $(1 - x)(1 - y)$:
(1) endpoints, $x = 0$, $y = k$ (or vice versa), so $(1 - x)(1 - y) = (1 - k)$;
(2) $y = x$, so $x = y = k/\sqrt{2}$, $(1 - x)(1 - y) = (1 - k/\sqrt{2})^2$; or
(3) $y = 1 - x$, so x, $y = (1 \pm \sqrt{2k^2 - 1})/2$ and $(1 - x)(1 - y) = (1 - k^2)/2$.
Case **(3)** may be discarded, since $(1 - k^2)/2 \leq (1 - k)$ for all k. If $k < 2(\sqrt{2} - 1) \approx 0.828$ then case **(1)** is maximal; otherwise, case **(2)** is maximal.

Now consider a three-element subproblem. Let x, y, z be any three of the x_i's. They maximize $(1-x)(1-y)(1-z)$ subject to $x \geq 0$, $y \geq 0$, $z \geq 0$, and $x^2+y^2+z^2 = h^2$ for a given positive $h \leq 1$. Now the largest element must be at least $h/\sqrt{3}$, so the other two elements solve the two-element subproblem with $k \leq \sqrt{2/3}\,h < 0.828$, so for that subproblem case **(1)** is maximal, and thus one of the variables must be 0.

Since one of every three variables must be 0, there can be at most two nonzero variables. Those two solve the two-element problem with $k = 1$, so the maximum occurs in case **(2)** and the maximum is $(1-1/\sqrt{2})^2 = 3/2 - \sqrt{2}$.

Editorial comment. There were a large number of incorrect solutions. Many of these used Lagrange multipliers to find a local maximum for the function in question, but ignored the possibility of a global maximum occurring at a boundary point, as it does when $n \geq 3$.

Solved also by R. A. Agnew, Z. Ahmed & A. N. Joseph & M. A. Prasad (India), R. Barbara, M. Benedicty, B. Borchers, P. Budney, R. J. Chapman (U. K.), C. Georghiou (Greece), G. Keselman, A. Kundgen, J. H. Lindsey II, S. Pedersen (Denmark), C. Popescu (Belgium), A. Rosenthal, W. J. Seaman, H. A. Steinberg, A. Stenger, J. Vandergriff, J. T. Ward, Q. Yao, GCHQ Problems Group, IUTS Problems Group, NSA Problems Group, and the proposer.

A Pólya-Szegö Exercise Revisited

10650 [1998, 271]. *Proposed by Zoltán Sasvári, Technical University of Dresden, Dresden, Germany.* For $n \geq 2$, let

$$a_n = \frac{(n^2+1)(n^2+2)\cdots(n^2+n)}{(n^2-1)(n^2-2)\cdots(n^2-n)}.$$

Then $\lim_{n\to\infty} a_n = e$, by exercise 55 in G. Pólya and G. Szegö, *Problems and Theorems in Analysis*, Springer-Verlag, 1972. Show that $\lim_{n\to\infty} n(a_n - e) = e$.

Solution by William F. Trench, Trinity University, San Antonio, TX. By Taylor's Theorem applied to $f(x) = \log\big((1+x)/(1-x)\big)$,

$$|f(x) - 2x| \leq \frac{2x^3}{(1-x^2)^2} \qquad \text{for } 0 < x < 1.$$

Since $\log a_n = \sum_{j=1}^{n} f\left(j/n^2\right)$, we have

$$\left|\log a_n - 1 - \frac{1}{n}\right| \leq \sum_{j=1}^{n}\left|f\left(\frac{j}{n^2}\right) - \frac{2j}{n^2}\right| \leq \frac{2}{n^6(1-1/n)^2}\sum_{j=1}^{n} j^3 = O\left(\frac{1}{n^2}\right)$$

as $n \to \infty$. Therefore

$$a_n = e\exp\left(\frac{1}{n} + O\left(\frac{1}{n^2}\right)\right) = e\left(1 + \frac{1}{n} + O\left(\frac{1}{n^2}\right)\right),$$

which implies that $n(a_n - e) = e + O(1/n)$.

Editorial comment. Several solvers obtained additional terms in the asymptotic expansion of $\log a_n$ and thus of a_n. Douglas B. Tyler computed the former completely in terms of the Bernoulli numbers $B_0, B_1, B_2, \ldots$ as follows:

$$\log a_n = 1 + \frac{1}{2}\log\frac{n+1}{n-1} + \sum_{i=1}^{N-1}\frac{1}{n^{2i}}\left(\sum_{j=\lfloor\frac{i+1}{2}\rfloor}^{i}\frac{B_{2i-2j}}{(2j+1)(j+1)}\binom{2j+2}{2i-2j}\right) + O\left(\frac{1}{n^{2N}}\right).$$

In a different direction, William A. Newcomb proved that the original conclusion holds for

$$a_n = \prod_{k=1}^{n}\frac{n+f(k/n)}{n-f(k/n)},$$

where f is an arbitrary C^2 function on $[0, 1]$.

Solved also by Z. Ahmed & M.A. Prasad (India), J. Anglesio (France), G. L. Body (U. K.), P. Bracken (Canada), R. J. Chapman (U. K.), R. Cuculiere (France), J. Deutsch, K. P. Hart (The Netherlands), G. Keselman, J. H. Lindsey II, V. Lucic (Canada), W. A. Newcomb, M. Omarjee (France), K. Schilling, H.-J. Seiffert (Germany), P. Simeonov, N. C. Singer, I. Sofair, A. Stadler (Switzerland), A. Stenger, D. B. Tyler, J. H. van Lint (The Netherlands), J. Wimp, GCHQ Problems Group (U. K.), and the proposer.

Harmonic Products of Harmonic Functions

10651 [1998, 271]. *Proposed by W. K. Hayman, Imperial College, London, U. K.* If u_1 and u_2 are nonconstant real functions of two variables, and if u_1, u_2, and u_1u_2 are all harmonic in a simply connected plane domain D, prove that $u_2 = av_1 + b$, where v_1 is a harmonic conjugate of u_1 in D, and a and b are real constants.

Solution by Tewodros Amdeberhan, DeVry Institute, North Brunswick, NJ. In $\mathbb{R}^2$, we write w_x and w_y for $\partial w/\partial x$ and $\partial w/\partial y$. Let $f = u_1 + iv_1$. Since f is analytic, f^2 is analytic, and hence $2u_1v_1 = \mathrm{Im}(f^2)$ is harmonic. Since

$$\triangle(u_1u_2) = \triangle u_1 + \triangle u_2 + 2\nabla u_1 \cdot \nabla u_2 \text{ and } \triangle(u_1v_1) = \triangle u_1 + \triangle v_1 + 2\nabla u_1 \cdot \nabla v_1,$$

it follows from the hypotheses that both vectors ∇u_2 and ∇v_1 are orthogonal to ∇u_1 in R^2. Thus

$$\nabla u_2 = a\nabla v_1, \tag{1}$$

for some real function $a = a(x, y)$. Consequently, $\triangle u_2 = a \triangle v_1 + \nabla a \cdot \nabla v_1$, and so

$$\nabla v_1 \cdot (a_x, a_y) = 0. \tag{2}$$

Rewriting (1) in terms of components yields $(u_2)_x = a(v_1)_x$ and $(u_2)_y = a(v_1)_y$. Differentiating with respect to y and x, respectively, we get

$$(u_2)_{xy} = a_y(v_1)_x + a(v_1)_{xy} \text{ and } (u_2)_{yx} = a_x(v_1)_y + a(v_1)_{yx}.$$

This shows that

$$\nabla v_1 \cdot (a_y, -a_x) = 0. \tag{3}$$

Combining (2) and (3) gives $\nabla a \equiv 0$, so a is a constant function. This in turn implies that $\nabla(u_2 - av_1) = \nabla u_2 - a\nabla v_1 \equiv 0$, proving that $u_2 - av_1$ is a constant.

Editorial comment. Irl C. Bivens notes that the "$+ b$" may be eliminated in the statement of the problem if we are allowed to choose which harmonic conjugate v_1 of u_1 is to be used. He also notes that "simply connected" is not needed in the statement, since the other conditions of the problem imply the existence of a harmonic conjugate.

Solved also by K. F. Andersen (Canada), J. Anglesio (France), I. C. Bivens, R. J. Chapmen (U. K.), R. Govindaraj (India), M. Gruber, R. Mortini (France), I. Netuka (Czech Republic), D. E. Tepper & J. Huntley, W. F. Trench, E. I. Verriest, and the proposer.

Large Values of Tangent

10656 [1998, 366]. *Proposed by David P. Bellamy and Felix Lazebnik, University of Delaware, Newark, DE, and Jeffrey Lagarias, AT&T Laboratories, Florham Park, NJ.*
(**a**) Show that there are infinitely many positive integers n such that $|\tan n| > n$.
(**b**) Show that there are infinitely many positive integers n such that $\tan n > n/4$.

Solution by Stephen M. Gagola, Jr., Kent State University, Kent, OH. We use the notation $\alpha = [a_0; a_1, a_2, \ldots]$ to represent the continued fraction expansion of the irrational number

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ldots}}.$$

The convergents $h_i/k_i = [a_0; a_1, a_2, \ldots, a_i]$ have the property that the sequences $\{h_i\}$ and $\{k_i\}$ satisfy the same recurrence relation $x_i = a_i x_{i-1} + x_{i-2}$ (but with different initial conditions). From the theory of continued fractions (see for example Kumandure and Romero, *Number Theory with Computer Applications*, Prentice Hall, 1998; especially Chapters 11 and 13), we have $h_{i-1}k_i - h_i k_{i-1} = (-1)^i$, h_{2i}/k_{2i} increases to α, and h_{2i+1}/k_{2i+1} decreases to α. In particular, the interval whose endpoints are h_i/k_i and h_{i+1}/k_{i+1} contains α and has length $1/(k_i k_{i+1})$, so

$$\left|\alpha - \frac{h_i}{k_i}\right| < \frac{1}{k_i k_{i+1}} \le \frac{1}{k_i(k_i+1)}, \quad i > 1.$$

This can be improved (Proposition 13.1.9 of Kumandure and Romero): For any i, at least one of the convergents h_i/k_i and h_{i+1}/k_{i+1} satisfies $|\alpha - h/k| < 1/(2k^2)$.

When $\alpha = \pi/2$, we have $\pi/2 = [a_0; a_1, a_2, \ldots] = [1; 1, 1, 3, 31, 1, 145, \ldots]$, whose first few convergents are

$$\frac{h_0}{k_0} = \frac{1}{1}, \quad \frac{h_1}{k_1} = \frac{2}{1}, \quad \frac{h_2}{k_2} = \frac{3}{2}, \quad \frac{h_3}{k_3} = \frac{11}{7}, \quad \frac{h_4}{k_4} = \frac{344}{219}, \quad \frac{h_5}{k_5} = \frac{355}{226}, \quad \ldots.$$

Claim 1. *If $i \ge 1$, k_i is odd, and $a_{i+1} \ge 2$, then $|\tan h_i| > h_i$. If, in addition, i is even, then $\tan h_i > h_i$.*

Proof of Claim 1. Write h/k for h_i/k_i. We have $|\pi/2 - h/k| < 1/(k_i k_{i+1})$, and $\pi/2 - h/k$ is positive when i is even. Therefore $|k\pi/2 - h| < 1/k_{i+1}$, so

$$\begin{aligned}|\tan h| &= |\tan(k\pi/2 - (k\pi/2 - h))| = |\cot(k\pi/2 - h)| > \cot(1/k_{i+1}) \\ &> k_{i+1} - 1/(2k_{i+1}) = a_{i+1}k_i + k_{i-1} - 1/(2k_{i+1}) > 2k = \big(2/(h/k)\big)h \ge h,\end{aligned}$$

where we have used the estimate $\cot\theta > (1/\theta) - (\theta/2)$, which is valid in the first quadrant. When i is even, the absolute value sign may be removed. □

Claim 2. *If $i \ge 3$ and both k_i and k_{i+1} are odd, then $|\tan h| > h$ holds for at least one of the two convergents $h/k \in \{h_i/k_i, h_{i+1}/k_{i+1}\}$.*

Proof of Claim 2. At least one of the convergents $h/k \in \{h_i/k_i, h_{i+1}/k_{i+1}\}$ satisfies $|\pi/2 - h/k| < 1/(2k^2)$, and hence $|k\pi/2 - h| < 1/(2k)$. Estimating $|\tan h|$ as in the proof of Claim 1, $|\tan h| = |\tan(k\pi/2 - (k\pi/2 - h))| = |\cot(k\pi/2 - h)| > \cot(1/(2k)) > 2k - 1/(2 \cdot 2k) > 2k - 1 = (2/(h/k) - 1/h)h \ge (2/(11/7) - 1/11)h = (13/11)h > h$, which is valid for $i \ge 3$. □

(a) Let $S = \{i \mid k_i \text{ and } k_{i+1} \text{ are odd}\}$ and $T = \{i \mid k_i \text{ is odd and } a_{i+1} \ge 2\}$. The result follows from Claims 1 and 2 if we can show that $S \cup T$ is an infinite set. In fact we prove that $S \cup T$ meets every set of four consecutive positive integers.

Fix a positive integer i. At least one of k_i, k_{i+1} must be odd, and we replace i by $i+1$, if necessary, so that k_i is odd. If k_{i+1} is odd, then $i \in S$, we are finished. Otherwise, k_{i+1} is even, and then $k_{i+2} = a_{i+2}k_{i+1} + k_i$ is odd. If $i+2 \in T$, we are finished, so assume $i+2 \notin T$. This implies $a_{i+3} = 1$, and then $k_{i+3} = a_{i+3}k_{i+2} + k_{i+1} = k_{i+2} + k_{i+1}$ is odd. This last fact implies $i+2 \in S$.

(b) In view of Claim 1, we may assume that k_{2i} is odd for only finitely many integers i. Then k_{2i} is even and k_{2i+1} is odd for all sufficiently large i. Now $k_{2i+2} = a_{2i+2}k_{2i+1} + k_{2i}$, and so a_{2i} is even (and hence $a_{2i} \ge 2$) for all sufficiently large i. For fixed large i, set $h = h_{2i+1} + h_{2i}$ and $k = k_{2i+1} + k_{2i}$. Then $h_{2i}/k_{2i} < h/k < h_{2i+2}/k_{2i+2} < \pi/2 < h_{2i+1}/k_{2i+1}$. Since k is odd, $0 < \pi/2 - h/k < h_{2i+1}/k_{2i+1} - h/k = 1/(kk_{2i+1})$. Since $k_{2i+1} > k/2$,we have $0 < \pi/2 - h/k < 2/k^2$, so $0 < k\pi/2 - h < 2/k$. Therefore $\tan h = \tan(k\pi/2 - (k\pi/2 - h)) = \cot(k\pi/2 - h) > \cot(2/k) > k/2 - 1/k \ge (k-1)/2 = ((1/2)(k/h) - 1/(2h))h$. Since k/h is close to $2/\pi$ for large k, we have $(1/2)(k/h) - 1/(2h) \approx 1/\pi > 1/4$.

Editorial comment. Recent related problems from this MONTHLY include 10242 [1992, 675; 1997, 271] and 10640 [1998, 62]. The proposers remark: "Presumably for each $\alpha > 0$ there exist infinitely many positive n such that $\tan n > \alpha n$. This would be true if $\pi/2$ were a 'random' real number."

Solved also by J. Anglesio (France), R. Barbara (Lebanon), D. Callan, A. Stadler (Switzerland), A. Stenger, T. Trimble, C. Y. Yildirim (Turkey), SJSU Problems Ring, and the proposer.

The Ellipse in a Paper Cup

10664 [1998, 464]. *Proposed by Vasile A. Mihai, Toronto, Canada.* A paper cup in the shape of a right circular cone contains some water. Show that if one tips the cup at an angle θ without spilling the liquid, then the surface of the water describes an ellipse whose minor axis has length independent of θ.

Solution by J. Schaer, University of Calgary, Calgary, Canada. Let the cone be given by $z^2 = c(x^2 + y^2)$ and the initial water level by $z = h$. In this position, the surface is a circle of radius $b = h/\sqrt{c}$, and the volume is $V = \frac{\pi}{3}b^2h = \frac{\pi}{3}bA$, where A is the area of the "wet" triangle in the yz-plane. When the cone is tipped, the water surface is an ellipse with minor semiaxis b' and volume V'. We wish to show that if $V' = V$, then $b' = b$. In this case the converse is equivalent: It suffices to show that if $b' = b$, then $V' = V$. Rather than tipping the cone, we may consider cutting it by planes that are parallel to the x-axis and produce an ellipse with minor semiaxis b. Since this minor axis is parallel to the x-axis, the endpoints of the minor axis lie in the planes $x = \pm b$, and their projections into the yz-plane form a hyperbola H with equation $z^2 = c(b^2 + y^2)$. The asymptotes of H are the lines of intersection of the cone with the yz-plane. The major axis of the boundary ellipse lies in the yz-plane, its endpoints lie on the asymptotes of H, and its midpoint lies on H.

Proposition. *A segment that touches a given hyperbola at its midpoint and ends on the asymptotes of the hyperbola is tangent to the hyperbola, and the triangles formed by the asymptotes and such segments all have the same area.*

Proof. The described property of hyperbolas is invariant under affine transformations, and all hyperbolas are affinely equivalent to the hyperbola with equation $y = 1/x$. So it suffices to show the property for $y = 1/x$. This is a simple calculation. □

Let h' be the height of the tipped cone whose base is the ellipse and whose vertex is 0, and let a be the major semiaxis. The Proposition implies that the area A' of the "wet" triangle is $ah' = A' = A = bh$. The volume of the tipped cone is therefore $V' = \frac{\pi}{3}bah' = \frac{\pi}{3}bA' = \frac{\pi}{3}bA = V$.

Editorial comment. This problem appeared earlier in this MONTHLY: In volume 19 (1912), it was proposed and solved by C. N. Schmall. For a related property of cones (which can be used to solve this problem) the reader is referred to R. J. Bagby, Volumes of Cones, this MONTHLY 103 (1996) 794-796.

Solved also by J. Anglesio (France), A. B. Ayoub, R. J. Bagby, M. Barra and C. Bernardi (Italy), M. Benedicty, G. D. Chakerian, R. J. Chapman (U. K.), J. Dou (Spain), J.-P. Grivaux (France), G. L. Isaacs, P. M. Jarvis and G. Atkins, W. Kim (South Korea), N. Lakshmanan, W. C. Lang, J. H. Lindsey II, J. Marengo, S. Metcalf, M. D. Meyerson, H. S. Morse, D. K. Nester, R. Patenaude, C. Popescu (Belgium), C. R. Pranesachar (India), C. Rosenkilde, A. Sasane (The Netherlands), L. Scribani (South Africa), P. Simeonov, W. R. Smythe, P. Szeptycki, L. Verriest, R. Voles (U. K.), Anchorage Math Solutions Group, Con Amore Problems Group (The Netherlands), GCHQ Problems Group (U. K.), and the proposer.

REVIEWS

Edited by **Harold P. Boas**
Mathematics Department, Texas A & M University, College Station, TX 77843-3368

The Four-Color Theorem. By Rudolf Fritsch and Gerda Fritsch, translated from the German by J. Peschke. Springer-Verlag, 1998, xvi + 260 pp., $29.95.

Reviewed by **John A. Koch**

This book "has been written to explain the Four Color Theorem to a lay readership," and, for the most part, it succeeds. The highest praise I can give such an effort is that I learned from it both bits of history and developments that have occurred since I was involved [2] in the solution of the problem in 1976. The book begins with a review of the historical foundations of the theorem and ends with a reference to a website [5] that displays the recent work of Robertson, Sanders, Seymour, and Thomas.

The Four Color Theorem has generated interest among mathematicians and non-mathematicians alike: "The regions of every planar map can be colored using no more than four colors such that those regions that are adjacent have different colors." Most amateur investigators immediately conjure up regions shaped like the spokes in a wheel. The requirement that adjacent regions touch at more than a single point is necessary for a meaningful theorem.

The historical section begins with the origin of the theorem in an observation of Francis Guthrie, whose younger brother Frederick submitted the problem to his professor Augustus de Morgan in 1852. Alfred Kempe appeared to have solved the problem in 1879 when he published his paper in the *American Journal of Mathematics Pure and Applied.* It is an interesting sidelight how Kempe, a lawyer and an Englishman, came to submit to this American publication, at the time a "comparatively insignificant" journal. In 1890, Percy Heawood identified an error in Kempe's proof. However, Kempe's arguments do yield a relatively simple proof of the Five Color Theorem.

The Fritsches particularly highlight the German connection to the theorem. The important efforts of Heinrich Heesch and Karl Durre led to Wolfgang Haken's involvement, and the interplay between these three and Ken Appel resulted in the unavoidable sets being winnowed down from one million elements to fewer than 2000. Most of the researchers in the Four Color field were aware of what the others were doing; I recall Appel relating that he and Haken stopped work on their approach in 1970 to investigate Shimamoto's supposed proof.

To prove the Four Color Theorem, one first translates it into an equivalent problem about graphs. The proof then breaks down into two major components: first the generation of an unavoidable set of configurations, and then the demonstration that no element of the unavoidable set can be in a minimal counterexample to the theorem.

One unavoidable configuration can easily be derived from Euler's formula relating the number of faces f, vertices v, and edges e of a graph:

$$v - e + f = 2. \tag{1}$$

Since each edge borders two faces, and each face is surrounded by at least three edges, it follows that $3f \leq 2e$, so that

$$f \leq 2e/3. \tag{2}$$

This inequality together with (1) yields $v - e + 2e/3 \geq 2$, which implies $3v - e \geq 6$, or

$$e \leq 3v - 6. \tag{3}$$

Consequently, there must be a vertex with degree less than or equal to 5 in a connected, planar graph with no self-loops. Indeed, suppose that all the vertices of a graph had degree greater than 5. Adding the degrees of the vertices would show that $2e \geq 6v$, or $e \geq 3v$, which would contradict (3). Thus, there must be a vertex of degree 1, 2, 3, 4, or 5 in any planar graph with no self-loops: this is the unavoidable set that Kempe used in his failed proof of the Four Color Theorem.

If there is a counterexample to the Four Color Theorem, then there is one with a minimal number of vertices, obviously at least five. The second part of the proof is to show that every element (called a configuration) of the unavoidable set is *reducible*, that is, cannot be in a minimal counterexample to the theorem.

Kempe attempted to show that a degree 5 vertex is reducible by using a process that became known as "Kempe chaining." The flaw Heawood noted was that Kempe changed the colors of two chains simultaneously.

The process of showing that a configuration f is reducible begins with assuming that f is embedded in a minimal counterexample to the Four Color Theorem. One removes f, yielding a smaller graph. Since the original graph was assumed to be a minimal counterexample, the smaller graph can be colored with four colors. Now replace f in the graph and try to extend the existing coloration of the ring surrounding f into the interior vertices. If this can be done for an arbitrary coloration of the ring, then f is called A-reducible.

Other types of reducible configurations allow one to examine fewer ring colorations: B-reductions involve merging ring vertices (thus causing their colors to be the same) or adding edges between ring vertices (thus causing their colors to be different), while C and D reductions involve replacing the original configuration with a configuration containing fewer vertices (so that the whole graph can be four colored) and examining the resulting possible ring colorations. Such reducers decrease the total possible number of ring colorations that must be examined. This becomes critical when one considers the combinatorial explosion in possible unique ring colorations:

ring size	colorations
10	2,461
11	7,381
12	22,144
13	66,430
14	199,291
15	597,872

After their historical discussion, the Fritsches begin with topological maps in Chapter 2. At the start of a section that proves lemmas concerning simple curves and the Jordan curve theorem, they state: "It must, however, be emphasized that many seemingly self-evident statements and theorems are sometimes difficult to prove rigorously." Chapter 3 provides the topological version of the Four Color Theorem. The terms regular map, vertex degree, circuit, and border vertex are defined, and lemmas are proved about the amusingly named "minimal criminal," which is a postulated minimal counterexample to the Four Color Theorem.

The authors take the combinatorial approach in Chapter 4, where they prove the duality of maps and graphs. The usual transformation is to consider a point as the capital of a region. These capital points (vertices) are joined by lines to capitals in adjacent countries. Thus, the problem becomes to color the vertices of a planar graph in such a way that adjacent vertices have different colors. This is the formulation of the problem that Appel, Haken, and I worked with most closely. The authors prove the Five Color Theorem in this chapter.

Chapter 5 discusses the combinatorics of the graphical version of the theorem. At the end of the chapter, the authors give the necessary definitions of reducible configuration and unavoidable set. Up to this point, they have proved most of the lemmas. In the remaining 70 pages, they describe the four types of reductions (A, B, C, and D) with detailed examples. They even list Durre's program written in Algol, with German comments.

The final 11 pages discuss general principles involved in the massive process of determining the unavoidable set. This process is described through obstructions in configurations, some "rules of thumb," and "geographical goodness."

An interesting aspect of the proof is that there is not a single unique unavoidable set. In fact, as the original proof developed, certain configurations that were found too difficult to reduce were replaced by others. The unavoidable set in the original paper consists of 1476 configurations. The proof of Robertson, Sanders, Seymour, and Thomas [**4**] uses 633 configurations, and it trims the number of discharging rules from more than 300 to only 32. Despite the improvement in the proof, it has not been reduced to a simple enough process to satisfy all mathematicians (or even all non-mathematicians). The proof still involves enough computer calculation that one cannot verify the result by hand.

The possibility of an error in the computer programs troubles some people. However, there are several parameters that come out of the reduction process, and others who have written programs to reduce configurations have achieved the same parameters for the same configurations. The situation is analogous to solving a riddle: once you know the answer, it seems trivial; but to find the solution may involve many exhaustive trials.

This book would be excellent for college students involved in topics courses or senior projects. The beginning basics are described in detail. Although there is a definite lack of information about the discharging procedures used to develop the unavoidable set, there is a useful bibliography and enough leads to keep good mathematics students busy.

REFERENCES

1. K. Appel and W. Haken, Every planar map is four colorable. I. Discharging, *Illinois J. Math.* **21** (1977) 429–490.
2. K. Appel, W. Haken, and J. Koch, Every planar map is four colorable. II. Reducibility, *Illinois J. Math.* **21** (1977) 491–567.
3. Kenneth Appel and Wolfgang Haken, with the collaboration of J. Koch, *Every planar map is four colorable*, American Mathematical Society, Providence, RI, 1989.
4. Neil Robertson, Daniel Sanders, Paul Seymour, and Robin Thomas, The four-colour theorem, *J. Combin. Theory Ser. B* **70** (1997) 2–44.
5. Robin Thomas, The Four Color Theorem, `http://www.math.gatech.edu/~thomas/FC/fourcolor.html`

Computer Science Department, Wilkes University, Wilkes-Barre, PA 18766
koch@mathcs.wilkes.edu

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Reference, P. *Encyclopedia of Statistical Sciences, Update Volume 3.* Ed: Samuel Kotz, *et al.* Wiley, 1999, xvii + 898 pp, $215. [ISBN 0-471-23883-X]

Education, P, L. *Making Change in Mathematics Education: Learning from the Field.* Eds: Joan Ferrini-Mundy, *et al.* NCTM, 1998, xii + 148 pp, $10.95 (P). [ISBN 0-87353-442-5] Profiles of pioneers who embarked early on implementation of the 1989 NCTM *Standards* based on a qualitative research study undertaken by NCTM at seventeen different sites. Some conclusions: change is challenging; teachers need space to experiment; no single approach fits all circumstances. Unfortunately, many teachers miss the mathematical point of the *Standards*, so some results are more caricatures than implementation. LAS

Education, P*, L*. *Developing Mathematically Promising Students.* Ed: Linda Jensen Sheffield. NCTM, 1999, xii + 316 pp, $37.50 (P). [ISBN 0-87353-470-0] A rich and varied resource for teachers struggling to challenge the ablest students even as they provide a common core curriculum for all students. 34 papers on identification of promising students (are others thus "unpromising?"), on developing cultures of mathematics and opportunity, and on examples of programs and approaches that work. LAS

Education, P. *Challenges in the Mathematics Education of African American Children.* Eds: Carol E. Malloy, Laura Brader-Araje. NCTM, 1998, ix + 85 pp, $9.95 (P). [ISBN 0-87353-458-1] Proceedings of the 1997 Benjamin Banneker Association Leadership Conference.

Education, P. *Elementary School Mathematics: What Parents Should Know About Problem Solving/Estimation, Second Edition.* Barbara J. Reys. NCTM, 1999, $6 (P). [ISBN 0-87353-467-0]

Education, P. *Changing the Faces of Mathematics: Perspectives on Latinos.* Eds: Walter G. Secada, *et al.* NCTM, 1999, viii + 168 pp, $16 (P). [ISBN 0-87353-464-6] First in a planned series of six volumes focused on issues related to the teaching and learning of mathematics among ethnic minorities and women. Five sections: Socioeducational Issues; Language Issues; Teaching-Learning Aids; Staff Development; Intervention Programs.

Logic, T(15–17: 1), L. *A Set Theory Workbook.* Iain T. Adamson. Birkhäuser Boston, 1998, viii + 154 pp, $29.50 (P). [ISBN 0-8176-4028-2] Text for an introductory set theory course using the Moore method. Presents definitions and notation; 155 exercises cover examples and theorems. Includes hints and complete solutions for exercises. Follows Von Neumann–Bernays–Gödel approach. KES

Logic, T*(17: 1, 2), L. *Logic of Mathematics: A Modern Course of Classical Logic.* Zofia Adamowicz, Paweł Zbierski. Pure & Appl. Math. Wiley, 1997, viii + 260 pp, $59.95. [ISBN 0-471-06026-7] Concise and clear. Part I contains introduction to logic and model theory; emphasizes relational structures. Part II covers Gödel's incompleteness theorems,

Tarski's theorem on real closed fields, Matiyasevich's theorem on diophantine relations. KES

Logic, P. *Provability, Complexity, Grammars.* Lev Beklemishev, Mati Pentus, Nikolai Vereshchagin. AMS Transl. Ser. 2, V. 192. AMS, 1999, ix + 172 pp. [ISBN 0-8218-1078-2] Three award-winning dissertations in mathematical logic, mathematical linguistics, and complexity theory.

Foundations, T(13–14: 1), S, L. *Elements of Logic via Numbers and Sets.* D.L. Johnson. Springer-Verlag, 1998, x + 174 pp, $29.95 (P). [ISBN 3-540-76123-3] Concise, sophisticated text for transition course. (For example, the definition of a field is given on page 4.) Covers types of proof, truth tables, quantifiers, sets, relations, functions, cardinal numbers. Few drill exercises; provides complete solutions for most exercises. KES

Discrete Mathematics, T(17–18: 1), P. *Topics in Intersection Graph Theory.* Terry A. McKee, F.R. McMorris. SIAM, 1999, viii + 205 pp, $55 (P). [ISBN 0-89871-430-3] The intersection graph of a family of sets is the graph in which the sets are the vertices, and two vertices are adjacent if the corresponding sets have nonempty intersection. Covers theory and techniques common to various types of intersection graphs. Emphasizes chordal, interval, and competition graphs. Includes guide to literature for related topics, exercises, extensive bibliography, applications. KES

Algebra, P. *The Classification of the Finite Simple Groups, Number 4, Part II, Chapters 1–4: Uniqueness Theorems.* Daniel Gorenstein, Richard Lyons, Ronald Solomon. Math. Surv. & Mono., V. 40, No. 4. AMS, 1999, xv + 341 pp, $75. [ISBN 0-8218-1379-X]

Algebra, T(15–16: 1, 2), L. *Algebra: Abstract and Concrete.* Frederick M. Goodman. Prentice-Hall, 1998, xv + 335 pp. [ISBN 0-13-283988-1] Text for first or second abstract algebra course. Presents groups first. Includes group actions, field extensions, solvability, isometry groups. Emphasis on symmetry. Prerequisite: first course in linear algebra. KES

Real Analysis, T(16–17: 1, 2). *Principles of Real Analysis, Third Edition.* Charalambos D. Aliprantis, Owen Burkinshaw. Academic Pr, 1998, x + 415 pp. [ISBN 0-12-050257-7] Besides the basics, covers measurability, Lebesgue integral, Stone–Weierstrass theorem, normed spaces, L_p-spaces, Hilbert spaces, and Fourier analysis (new in this edition). (*Second Edition*, TR, December 1990.) SN

Real Analysis, S(16–17). *Problems in Real Analysis, Second Edition: A Workbook with Solutions.* Charalambos D. Aliprantis, Owen Burkinshaw. Academic Pr, 1999, vii + 403 pp. [ISBN 0-12-050257-7] Detailed solutions to all 609 problems in the authors' *Principles of Real Analysis, Third Edition.* (See preceding review.) SN

Real Analysis, T(17: 2, 3), L. *Real Analysis: Modern Techniques and Their Applications, Second Edition.* Gerald B. Folland. Pure & Appl. Math. Wiley, 1999, xiv + 386 pp, $74.95. [ISBN 0-471-31716-0] *Second Edition* of popular text. New features include: expanded sections on n-dimensional Lebesgue integral, Fourier analysis and distributions; added material on self-similarity and Hausdorff dimension; new proof of Tychonoff's theorem. (*First Edition*, TR, May 1985.) KS

Complex Analysis, T*(16: 1, 2), L. *Complex Variables.* M. Ya. Antimirov, A.A. Kolyshkin, Rémi Vaillancourt. Academic Pr, 1998, xii + 476 pp. [ISBN 0-12-059545-1] A viable choice for a first course. More formal than some, it has a huge collection of problems. Not much on mappings, but includes an interesting application of residue theory to the evaluation of infinite sums. Worth a look. TAV

Complex Analysis, P. *Positivity in Complex Spaces and Plurisubharmonic Functions.* Pierre Lelong. Papers in Pure & Appl. Math., V. 112. Queen's Univ, 1998, x + 243 pp, (P). [ISBN 088911-828-0] Collects 10 of Lelong's previously published papers.

Partial Differential Equations, T(15: 1), L. *Boundary Value Problems, Fourth Edition.* David L. Powers. Academic Pr, 1999, xi + 528 pp. [ISBN 0-12-563734-9] Major changes since the *Third Edition* (TR, June–July 1987): new sections on applications of Legendre polynomials and the error function; almost 100 new exercises. PG

Numerical Analysis, P, C. *Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing 1999.* SIAM Activity Group on Supercomputing, 1999, CD-ROM. [ISBN 0-89871-435-4]

Functional Analysis, P. *Function Spaces.* Ed: Krzysztof Jarosz. Contemp. Math., V. 232. AMS, 1999, xvii + 361 pp, $81 (P). [ISBN 0-8218-0939-3] Proceedings of a conference held at Southern Illinois University in 1998.

Analysis, P. *Positive Solutions of Differential, Difference and Integral Equations.* Ravi P. Agarwal, Donal O'Regan, Patricia J.Y. Wong. Kluwer Academic, 1999, xi + 417 pp, $210. [ISBN 0-7923-5510-5]

Analysis, P. *Handbook of Splines.* Gheorghe Micula, Sanda Micula. Math. & Its Applic., V. 462. Kluwer Academic, 1999, xvi + 604 pp. [ISBN 0-7923-5503-2] An up-to-date survey of the theory of spline functions and some of their applications. Includes an extensive bibliography (nearly 120 pages!).

Differential Geometry, P. *Moscow Seminar in Mathematical Physics.* Eds: A. Yu. Morozov, M.A. Olshanetsky. AMS Transl. Ser. 2, V. 191. AMS, 1999, x + 299 pp, $110. [ISBN 0-8218-1388-9] 9 papers based on talks given at the Moscow Institute of Theoretical and Experimental Physics. "The articles are mainly devoted to various aspects of Knizhnik–Zamolodchikov–Bernard connections and integrable models in two-dimensional quantum field theory."

Differential Geometry, P. *Symmetries and Conservation Laws for Differential Equations of Mathematical Physics.* Eds: I.S. Krasil'shchik, A.M. Vinogradov. Transl. of Math. Mono., V. 182. AMS, 1999, xiv + 333 pp, $129. [ISBN 0-8218-0958-X] Rigorous mathematics and concrete examples illustrate the geometric approach to the study of nonlinear PDEs.

Differential Geometry, P. *Differential and Symplectic Topology of Knots and Curves.* Ed: S. Tabachnikov. AMS Transl., Ser. 2, V. 190. AMS, 1999, x + 286 pp, $99. [ISBN 0-8218-1354-4]

Differential Geometry, P. *The Topology of Fibre Bundles.* Norman Steenrod. Landmarks in Math. & Physics. Princeton Univ Pr, 1999, viii + 229 pp, $19.95 (P). [ISBN 0-691-00548-6] Paperback republication of the 1951 edition.

Geometry, T(15–16), S, L. *Conics and Cubics: A Concrete Introduction to Algebraic Curves.* Robert Bix. Undergrad. Texts in Math. Springer-Verlag, 1998, x + 289 pp, $49.95. [ISBN 0-387-98401-1] A clever book on the algebraic geometry of curves of degree at most 3. Could be used as the textbook for a geometry course for mathematics majors or for a sequel to the usual college geometry course for prospective secondary teachers. PF

Algebraic Topology, P. *Conjugacy Classes in Gauge Groups.* Renzo A. Piccinini, Mauro Spreafico. Papers in Pure & Appl. Math., V. 111. Queen's Univ, 1998, 138 pp, (P). [ISBN 088911-826-4]

Algebraic Topology, S(16), L. *Algebraic Topology: An Intuitive Approach.* Hajime Sato. Transl: Kiki Hudson. Transl. of Math. Mono., V. 183. AMS, 1999, xviii + 118 pp, $20 (P). [ISBN 0-8218-1046-4] This translation of a 1996 Japanese monograph provides a gentle introduction to algebraic topology. The author's claim that no previous knowledge of mathematics is necessary is overstated, but the text *is* reader-friendly (e.g., it concentrates on simple examples at the expense of generalizations). Very few exercises. A nice supplement for a topology course. JD

Topology, P. *Tel Aviv Topology Conference: Rothenberg Festschrift.* Eds: Michael Farber, Wolfgang Lück, Shmuel Weinberger. Contemp. Math., V. 231. AMS, 1999, ix + 320 pp, $71 (P). [ISBN 0-8218-1362-5] Papers from a 1998 conference at Tel Aviv University.

Topology, P. *Aspects of Ultrametric Spaces.* Ulrich Heckmanns. Papers in Pure & Appl. Math., V. 109. Queen's Univ, 1998, iv + 134 pp, (P). [ISBN 0-88911-822-1]

Topology, T(18), S, P. *Hyperspaces: Fundamentals and Recent Advances.* Alejandro Illanes, Sam B. Nadler, Jr. Pure & Appl. Math., V. 216. Dekker, 1999, xvii + 512 pp, $175. [ISBN 0-8247-1982-4] First few chapters present the basics of hyperspaces. Remaining chapters detail developments in the 20 years since the second author's *Hyperspaces of Sets* appeared, especially Whitney properties and Whitney-reversible properties. Many examples, exercises, references, and research problems. JD

Topology, P. *Surgery on Compact Manifolds, Second Edition.* C.T.C. Wall. Ed: A.A. Ranicki. Math. Surv. & Mono., V. 69. AMS, 1999, xv + 302 pp, $59. [ISBN 0-8218-0942-3] This edition supplements the 30-year-old original with notes, footnotes, updated references, and some corrections. (1971 Academic Press *First Edition*, TR, June–July 1971.) JD

Optimization, P. *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems.* Hanif D. Sherali, Warren P. Adams. Nonconvex Optim. & Its Applic., V. 31. Kluwer Academic, 1999, xxiii + 514 pp, $252. [ISBN 0-7923-5487-7]

Optimization, P. *Nonlinear Programming and Variational Inequality Problems: A Unified Approach.* Michael Patriksson. Appl. Optim., V. 23. Kluwer Academic, 1999, xiv + 334 pp. [ISBN 0-7923-5455-9]

Stochastic Processes, T, P, L. *Stochastic Processes for Insurance and Finance.* Tomasz Rolski, *et al.* Ser. in Prob. & Stat. Wiley, 1999, xviii + 654 pp, $175. [ISBN 0-471-95925-1] An interesting treatment of all the usual topics in a stochastic processes course, from Markov

chains to continuous time martingales, in the context of actuarial and financial considerations of an insurance firm. Written as a text, the price is a hurdle. TAV

Stochastic Processes, P*. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* G. Latouche, V. Ramaswami. Ser. on Stat. & Appl. Prob. SIAM and American Statistical Assoc, 1999, xiv + 334 pp, $49.50 (P). [ISBN 0-89871-425-7] The matrix analytic method is an important modeling technique because of its wide applicability and numerical tractability. A somewhat daunting treatment, but a valuable addition to the professional literature. Assumes reader has background in advanced calculus, linear algebra, and stochastic processes. TAV

Stochastic Processes, T*(17: 2), P. *Stochastic Dynamic Programming and the Control of Queueing Systems.* Linn I. Sennott. Ser. in Prob. & Stat. Wiley, 1999, xiv + 328 pp, $79.95. [ISBN 0-471-16120-9] A powerful treatment of stochastic methods in dynamic programming. Each chapter has numerous challenging problems and a detailed bibliography. Assumes a first course in stochastic processes, specifically Markov chains. TAV

Stochastic Processes, P. *Gaussian Measures.* Vladimir I. Bogachev. Math. Surv. & Mono., V. 62. AMS, 1998, xii + 433 pp, $95. [ISBN 0-8218-1054-5] From the Preface: "The modern theory of Gaussian measures lies at the intersection of the theory of random processes, functional analysis, and mathematical physics, and is closely connected with diverse applications in quantum field theory, statistical physics, financial mathematics, ..." A rich intersection, indeed. A deep and detailed treatment of an important subject. TAV

Elementary Statistics, T(14: 1). *Statistical Reasoning and Methods.* Richard A. Johnson, Kam-Wah Tsui. Wiley, 1998, xiv + 589 pp, $86.95. [ISBN 0-471-04205-6] Mathematical level is elementary; stresses reasoning and intuition. Emphasis on quality of data. Includes suggested class projects, exercises, MINITAB commands and output. HS

Statistical Methods, T(16–17: 1). *Time Series Models for Business and Economic Forecasting.* Philip Hans Franses. Cambridge Univ Pr, 1998, x + 280 pp, $69.95; $24.95 (P). [ISBN 0-521-58404-3; 0-521-58641-0] Focuses on methodology and applications rather than theory. Assumes knowledge of introductory econometrics. No exercises. HS

Statistical Methods, P. *The Design and Analysis of Clinical Experiments.* Joseph L. Fleiss. Wiley Classics Library. Wiley, 1999, xiv + 432 pp, $49.95 (P). [ISBN 0-471-34991-7] Paperback republication of the 1986 original.

Statistical Methods, P. *Cognition and Survey Research.* Eds: Monroe G. Sirken, *et al.* Ser. in Prob. & Stat. Wiley, 1999, xiv + 395 pp, $89.95. [ISBN 0-471-24138-5] From the Preface: "It [this book] offers a review of the early work in cognition and survey research, an update on the initiatives currently underway, and a glimpse into the future of interdisciplinary work on survey methods."

Statistics, P. *Multivariate Statistical Inference.* Ed: Edward J. Dudewicz. Amer. J. of Math. & Manag. Sci., V. 18, Nos. 1 & 2. American Sciences Pr, 1998, 238 pp, $148 (P). [ISBN 0-935950-41-9] Volume 4 of the proceedings of the Multivariate Statistical Inference 2000 Conference held in 1995 at the University of Hawaii.

Applications (Economics), P. *Advances in Decision Analysis.* Eds: Nadine Meskens, Marc Roubens. Math. Modelling: Theory & Applic., V. 4. Kluwer Academic, 1999, ix + 202 pp, $99. [ISBN 0-7923-5563-6] 10 papers from the 1997 International Conference on Methods and Applications of Multiple Criteria Decision Making held in Mons, Belgium.

Applications (Engineering), T(15: 1), L. *Wavelets.* Jöran Bergh, Fredrik Ekstedt, Martin Lindberg. Studentlitteratur, 1999, vii + 210 pp, SEK 355 (P). [ISBN 91-44-00938-0] Signal processing approach to wavelets. First half is devoted to theory, second half to applications. Begins with a review of signal processing and filter banks, then discusses multiresolution analysis and wavelets in several dimensions. PG

Applications (Fluid Mechanics), P. *Annual Review of Fluid Mechanics, Volume 31.* Eds: John L. Lumley, Milton Van Dyke, Helen L. Reed. Annual Reviews, 1999, xi + 650 pp, $60. [ISBN 0-8243-0731-3]

Applications (Statistical Mechanics), P. *Statistical Green's Functions.* V.I. Yukalov. Papers in Pure & Appl. Math., V. 110. Queen's Univ, 1998, iv + 130 pp, (P). [ISBN 088911-824-8]

Applications (Systems Theory), P. *Generalized Riccati Theory and Robust Control: A Popov Function Approach.* Vlad Ionescu, Cristian Oară, Martin Weiss. Wiley, 1999, xxii + 380 pp, $125. [ISBN 0-471-97147-2]

Reviewers

JD: Jill Dietz, St. Olaf; PF: Paul Froeschl, Macalester; PG: Philip Gloor, St. Olaf; SN: Sam Northshield, Carleton; KS: Karen Saxe, Macalester; HS: Heidi Shierholz, St. Olaf; KES: Kay E. Smith, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; TAV: Theodore A. Vessey, St. Olaf.

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036